Finite Elements III: First-Order and Time-Dependent PDEs

Alexandre Ern, Jean-Luc Guermond

Finite Elements III:
First-Order and Time-Dependent PDEs

Alexandre Ern Jean-Luc Guermond

May 13, 2021

Contents

Part XII. First-order PDEs

Part XIII. Parabolic PDEs

Part XIV. Time-dependent Stokes equations

Part XV. Time-dependent first-order linear PDEs

Part XVI. Nonlinear hyperbolic PDEs

Chapter 56

Friedrichs' systems

In Part XII, composed of Chapters 56 to 63, we study the finite element approximation of PDEs where a coercivity property is not available, so that the analysis solely relies on inf-sup conditions. Stability can be obtained by employing various stabilization techniques (residual-based or fluctuation-based). In the present chapter, we introduce the prototypical model problem we are going to work on: it is a system of first-order linear PDEs introduced in 1958 by Friedrichs [131]. This system enjoys symmetry and positivity properties and is often referred to in the literature as *Friedrichs' system*. Friedrichs wanted to handle within a single functional framework PDEs that are partly elliptic and partly hyperbolic, and for this purpose he developed a formalism that goes beyond the traditional classification of PDEs into elliptic, parabolic, and hyperbolic types. Friedrichs' formalism is very powerful and encompasses several model problems. Important examples are the advection-reaction equation, the div-grad problem related to Darcy's equations, and the curl-curl problem related to Maxwell's equations. This theory will be used systematically in the following chapters. All the theoretical arguments in this chapter are presented assuming that the functions are complex-valued. The real-valued case can be obtained by replacing the field $\mathbb{C}$ by $\mathbb{R}$, by replacing the Hermitian transpose Z^{H} by the transpose Z^{T}, and by removing the real part symbol $\Re$.

56.1 Basic ideas

Let D be a Lipschitz domain in $\mathbb{R}^d$. We consider functions defined on D with values in $\mathbb{C}^m$ for some integer $m \geq 1$. The (Hermitian) inner product in $L := L^2(D;\mathbb{C}^m)$ is denoted by $(f,g)_L := \int_D g^{\mathsf{H}} f \,\mathrm{d}x$. Notice that $(f,g)_L = \overline{(g,f)_L}$ for all $f,g \in L$. Given two Hermitian matrices $\mathcal{B},\mathcal{C} \in \mathbb{C}^{m\times m}$ (i.e., $\mathcal{B} = \mathcal{B}^{\mathsf{H}}$, $\mathcal{C} = \mathcal{C}^{\mathsf{H}}$), the inequality $\mathcal{B} \geq \mathcal{C}$ means that $X^{\mathsf{H}}\mathcal{B}X \geq X^{\mathsf{H}}\mathcal{C}X$ for all $X \in \mathbb{C}^m$. We denote by $\mathbb{I}_m$ the identity matrix in $\mathbb{C}^{m\times m}$.

56.1.1 The fields $\mathcal{K}$ and $\mathcal{A}^k$

Let $\mathcal{K}$, $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ be a family of $(d+1)$ fields in $L^\infty(D;\mathbb{C}^{m\times m})$. We set $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k \mathcal{A}^k$ where $\partial_k := \frac{\partial}{\partial x_k}$. We make the following key assumptions:

$$\text{Boundedness:} \quad \mathcal{K},\ \{\mathcal{A}^k\}_{k\in\{1:d\}},\ \text{and } \mathcal{X} \text{ are in } L^\infty(D;\mathbb{C}^{m\times m}), \tag{56.1a}$$

$$\text{Symmetry:} \quad \mathcal{A}^k = (\mathcal{A}^k)^{\mathsf{H}} \text{ for all } k\in\{1:d\}, \text{ a.e. in } D, \tag{56.1b}$$

$$\text{Positivity:} \quad \exists \mu_0 > 0 \text{ s.t. } \mathcal{K} + \mathcal{K}^{\mathsf{H}} - \mathcal{X} \geq 2\mu_0 \mathbb{I}_m \text{ a.e. in } D. \tag{56.1c}$$

Notice that $\mathcal{X} = \mathcal{X}^{\mathsf{H}}$ owing to (56.1b). Using the above fields, it is possible to define the following differential operators on $C^1(\overline{D};\mathbb{C}^m)$:

$$A(v) := \mathcal{K}v + A_1(v), \qquad A_1(v) := \sum_{k\in\{1:d\}} \mathcal{A}^k \partial_k v. \tag{56.2}$$

56.1.2 Integration by parts

Let us assume for the time being that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are smooth enough to admit a bounded trace at the boundary ∂D. Let $(n_k)_{k\in\{1:d\}}$ be the Cartesian components of the outward unit normal $\boldsymbol{n}$. We define the boundary field $\mathcal{N} \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ by setting

$$\mathcal{N} := \sum_{k\in\{1:d\}} n_k \mathcal{A}^k_{|\partial D}. \tag{56.3}$$

Notice that $\mathcal{N} = \mathcal{N}^{\mathsf{H}}$ owing to (56.1b). Integration by parts is a key tool in the analysis of Friedrichs' systems. It involves in particular the formal adjoint $\tilde{A}$ of A, which is defined as follows: For all $v \in C^1(\overline{D};\mathbb{C}^m)$,

$$\tilde{A}(v) := (\mathcal{K}^{\mathsf{H}} - \mathcal{X})v - A_1(v) = (\mathcal{K} + \mathcal{K}^{\mathsf{H}} - \mathcal{X})v - A(v). \tag{56.4}$$

Lemma 56.1 (Integration by parts). *Let $L(\partial D) := L^2(\partial D;\mathbb{C}^m)$. The following holds true for all $v, w \in C^1(\overline{D};\mathbb{C}^m)$:*

$$(A(v), w)_L = (v, \tilde{A}(w))_L + (\mathcal{N}v, w)_{L(\partial D)}. \tag{56.5}$$

Proof. Using (56.1b) and the divergence formula, we infer that

$$\begin{aligned}
&(\mathcal{X}v, w)_L + (A_1(v), w)_L + (v, A_1(w))_L \\
&\qquad = \int_D \sum_{k\in\{1:d\}} \big(w^{\mathsf{H}}(\partial_k \mathcal{A}^k)v + w^{\mathsf{H}}\mathcal{A}^k\partial_k v + (\mathcal{A}^k\partial_k w)^{\mathsf{H}} v\big)\,\mathrm{d}x \\
&\qquad = \int_D \sum_{k\in\{1:d\}} \partial_k(w^{\mathsf{H}}\mathcal{A}^k v)\,\mathrm{d}x = \int_{\partial D} w^{\mathsf{H}}\mathcal{N}v\,\mathrm{d}s = (\mathcal{N}v, w)_{L(\partial D)}.
\end{aligned}$$

Since the field $\mathcal{X}$ takes Hermitian values, using (56.4) we then infer that

$$\begin{aligned}
(\mathcal{N}v, w)_{L(\partial D)} &= (A(v), w)_L - (\mathcal{K}v, w)_L + (\mathcal{X}v, w)_L + (v, A_1(w))_L \\
&= (A(v), w)_L - (v, \mathcal{K}^{\mathsf{H}}w)_L + (v, \mathcal{X}w)_L + (v, A_1(w))_L \\
&= (A(v), w)_L - (v, \tilde{A}(w))_L.
\end{aligned}$$

□

Lemma 56.2 (L-norm bound). *For all $v \in C^1(\overline{D};\mathbb{C}^m)$, we have*

$$\Re\big((A(v),v)_L\big) \geq \mu_0\|v\|_L^2 + \frac{1}{2}(\mathcal{N}v,v)_{L(\partial D)}. \tag{56.6}$$

Proof. Using (56.4) and Lemma 56.1, we infer that

$$\begin{aligned}\frac{1}{2}(A(v),v)_L &= \frac{1}{2}(v,\tilde{A}(v))_L + \frac{1}{2}(\mathcal{N}v,v)_{L(\partial D)}\\ &= -\frac{1}{2}\overline{(A(v),v)_L} + \frac{1}{2}((\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})v,v)_L + \frac{1}{2}(\mathcal{N}v,v)_{L(\partial D)},\end{aligned}$$

since $\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X}$ is Hermitian. This implies that $\Re\left((A(v),v)_L\right) = \frac{1}{2}((\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})v,v)_L + \frac{1}{2}(\mathcal{N}v,v)_{L(\partial D)}$, and (56.6) follows from (56.1c). □

The estimate (56.6) says that the sesquilinear form $(A(v),w)_L$ is L-coercive up to a boundary term. The key idea of Friedrichs is to enforce a suitable boundary condition to gain positivity on the boundary term. This is done by assuming that there exists another boundary field $\mathcal{M} \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ satisfying the following two algebraic properties a.e. on ∂D:

$$\mathcal{M} \text{ is nonnegative: } \Re(\xi^{\mathsf{H}}\mathcal{M}\xi) \geq 0 \text{ for all } \xi\in\mathbb{C}^m, \tag{56.7a}$$

$$\ker(\mathcal{M}-\mathcal{N}) + \ker(\mathcal{M}+\mathcal{N}) = \mathbb{C}^m. \tag{56.7b}$$

Note that the field $\mathcal{M}$ is not assumed to take Hermitian values. Since any function v satisfying $(\mathcal{M}-\mathcal{N})v_{|\partial D} = 0$ verifies $(\mathcal{M}v,v)_{L(\partial D)} \in \mathbb{R}$ (because $\mathcal{N}$ is Hermitian), we infer using (56.7a) in (56.6) that

$$\Re\left((A(v),v)_L\right) \geq \mu_0\|v\|_L^2 + \frac{1}{2}(\mathcal{M}v,v)_{L(\partial D)} \geq \mu_0\|v\|_L^2, \tag{56.8}$$

for every $v \in C^1(\overline{D};\mathbb{C}^m)$ such that $(\mathcal{M}-\mathcal{N})v_{|\partial D} = 0$.

56.1.3 The model problem

Given $f \in L$, our goal is to find a function $u : D \to \mathbb{C}^m$ such that

$$A(u) = f \ \text{ in } D, \qquad (\mathcal{M}-\mathcal{N})u = 0 \ \text{ on } \partial D. \tag{56.9}$$

Under the assumptions (56.1) and (56.7), Friedrichs proved: (i) the uniqueness of the strong solution $u \in C^1(\overline{D};\mathbb{C}^m)$ satisfying $(A(u),v)_L = (f,v)_L$ for all $v \in L$ and $(\mathcal{M}-\mathcal{N})u = 0$ on ∂D; (ii) the existence of an ultraweak solution $u \in L$ such that $(u,\tilde{A}(v))_L = (f,v)_L$ for all $v \in C^1(\overline{D};\mathbb{C}^m)$ such that $(\mathcal{M}^{\mathsf{H}}+\mathcal{N})v = 0$ on ∂D. In §56.3, we introduce a mathematical setting relying on boundary operators instead of boundary fields to define a notion of weak solution for (56.9), and we prove the well-posedness of the said formulation by using the BNB theorem.

56.2 Examples

This section presents three examples of Friedrichs' systems: the advection-reaction equation, Darcy's equations in mixed form, and the time-harmonic Maxwell's equations also in mixed form. These equations are written in dimensional form, and we refer the reader to §57.3.3 for a discussion on the rescaling of the various components of the unknown field u.

56.2.1 Advection-reaction equation

Let $\mu \in L^\infty(D;\mathbb{R})$ and let $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(D;\mathbb{R}^d)$ be such that $\nabla\cdot\boldsymbol{\beta} \in L^\infty(D;\mathbb{R})$. Notice that we work with $\mathbb{R}$-valued functions. Given $f \in L := L^2(D;\mathbb{R})$, we want to find a function $u : D \to \mathbb{R}$ such that

$$\mu u + \boldsymbol{\beta}\cdot\nabla u = f \quad \text{in } D. \tag{56.10}$$

This equation models the transport of a solute of concentration u by a flow field with velocity $\boldsymbol{\beta}$, reaction coefficient μ ($\mu \geq 0$ corresponds to depletion), and source term f. Typical SI units are s^{-1} for μ and $\mathrm{m}\cdot\mathrm{s}^{-1}$ for $\boldsymbol{\beta}$.

To recover Friedrichs' formalism, we set $m := 1$, $\mathcal{K} := \mu$, and $\mathcal{A}^k := \beta_k$ for all $k \in \{1{:}d\}$, where $(\beta_k)_{k\in\{1:d\}}$ denote the Cartesian components of $\boldsymbol{\beta}$. The assumption (56.1a) is satisfied since $\mu \in L^\infty(D)$, $\beta_k \in L^\infty(D)$ for all $k \in \{1{:}d\}$, and $\mathcal{X} = \nabla\cdot\boldsymbol{\beta} \in L^\infty(D)$. The assumption (56.1b) is trivially satisfied since $m = 1$. Finally, the assumption (56.1c) is satisfied provided we suppose that

$$\mu_0 := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D}(\mu - \tfrac{1}{2}\nabla\cdot\boldsymbol{\beta})(\boldsymbol{x}) > 0. \tag{56.11}$$

The boundary field is $\mathcal{N} := \boldsymbol{\beta}\cdot\boldsymbol{n}$, and the integration by parts formula (56.5) follows from the Leibniz product rule and the divergence formula, i.e.,

$$\int_D \big((\nabla\cdot\boldsymbol{\beta})vw + v(\boldsymbol{\beta}\cdot\nabla w) + w(\boldsymbol{\beta}\cdot\nabla v)\big)\,\mathrm{d}x = \int_D \nabla\cdot(\boldsymbol{\beta}vw)\,\mathrm{d}x = \int_{\partial D}(\boldsymbol{\beta}\cdot\boldsymbol{n})vw\,\mathrm{d}s.$$

To enforce a suitable boundary condition, we need to consider the sign of the normal component $\boldsymbol{\beta}\cdot\boldsymbol{n}$ at the boundary (see Figure 56.1). We define

$$\partial D^- := \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) < 0\}, \tag{56.12a}$$

$$\partial D^+ := \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) > 0\}, \tag{56.12b}$$

$$\partial D^0 := \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) = 0\}. \tag{56.12c}$$

Notice that both ∂D^+ and ∂D^- can be empty (think of a vector field $\boldsymbol{\beta}$ tangential to ∂D). We impose the *inflow boundary condition*

$$u = 0 \quad \text{on } \partial D^-. \tag{56.13}$$

This condition can be enforced by using the boundary field $\mathcal{M} := |\boldsymbol{\beta}\cdot\boldsymbol{n}|$. Indeed, $(\mathcal{M} - \mathcal{N})u = 0$ amounts to $(|\boldsymbol{\beta}\cdot\boldsymbol{n}| - \boldsymbol{\beta}\cdot\boldsymbol{n})u = 0$, i.e., $u = 0$ on ∂D^-. Notice that $\mathcal{M}$ satisfies (56.7a) trivially. Moreover, $\ker(\mathcal{M}(\boldsymbol{x}) - \mathcal{N}(\boldsymbol{x})) = \mathbb{R}$ and $\ker(\mathcal{M}(\boldsymbol{x}) + \mathcal{N}(\boldsymbol{x})) = \{0\}$ for a.e. $\boldsymbol{x} \in \partial D^+$, $\ker(\mathcal{M}(\boldsymbol{x}) - \mathcal{N}(\boldsymbol{x})) = \{0\}$ and $\ker(\mathcal{M}(\boldsymbol{x}) + \mathcal{N}(\boldsymbol{x})) = \mathbb{R}$ for a.e. $\boldsymbol{x} \in \partial D^-$, and $\ker(\mathcal{M}(\boldsymbol{x}) - \mathcal{N}(\boldsymbol{x})) = \ker(\mathcal{M}(\boldsymbol{x}) + \mathcal{N}(\boldsymbol{x})) = \mathbb{R}$ for a.e. $\boldsymbol{x} \in \partial D^0$, i.e., (56.7b) is satisfied in all the cases. In conclusion, $\mathcal{M}$ satisfies (56.7). Finally, for all $v \in C^1(\overline{D};\mathbb{C}^m)$ such that $(\mathcal{M} - \mathcal{N})v_{|\partial D} = 0$, the L-coercivity property (56.8) becomes

$$(A(v), v)_{L^2(D)} \geq \mu_0\|v\|^2_{L^2(D)} + \frac{1}{2}\int_{\partial D}|\boldsymbol{\beta}\cdot\boldsymbol{n}|v^2\,\mathrm{d}s. \tag{56.14}$$

Remark 56.3 (Hypothesis (56.11)). The hypothesis (56.11) is not satisfied if $\mu = 0$ and $\nabla\cdot\boldsymbol{\beta} = 0$. A well-posed weak formulation can still be derived if $\boldsymbol{\beta}$ is a filling field, i.e., if for a.e. $\boldsymbol{x} \in D$, there is a characteristic line of $\boldsymbol{\beta}$ that starts from ∂D^- and reaches $\boldsymbol{x}$ in finite time. More precisely, a sufficient condition is that there is a function $\zeta \in W^{1,\infty}(D)$ such that $\boldsymbol{\beta}\cdot\nabla\zeta$ is uniformly bounded away from zero; see §61.4. A simple example is the one-dimensional transport equation $u' = f$ in $D := (0,1)$ with $u(0) = 0$, i.e., $\boldsymbol{\beta} := \boldsymbol{e}_x$ and $\zeta := x$ (for instance); see §24.2.2. □

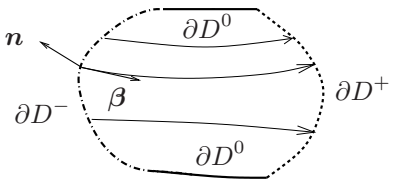

Figure 56.1: Advection-reaction problem: *inflow boundary* ∂D^-, *outflow boundary* ∂D^+, and *characteristic boundary* ∂D^0.

56.2.2 Darcy's equations

Let $\mu \in L^\infty(D;\mathbb{R})$ and let $\mathbb{d} \in \mathbb{L}^\infty(D) := L^\infty(D;\mathbb{R}^{d\times d})$ take symmetric values with eigenvalues in the interval $[\lambda_\flat(\boldsymbol{x}), \lambda_\sharp(\boldsymbol{x})]$ for a.e. $\boldsymbol{x} \in D$. Set $\lambda_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \lambda_\flat(\boldsymbol{x})$ and $\lambda_\sharp := \|\lambda_\sharp\|_{L^\infty(D)}$, and assume that $\lambda_\flat > 0$ (notice that $\lambda_\sharp$ is finite since $\mathbb{d} \in \mathbb{L}^\infty(D)$). Given $f \in L^2(D)$, we want to find a field $\boldsymbol{\sigma} : D \to \mathbb{R}^d$ and a function $p : D \to \mathbb{R}$ (notice that here again we work with real-valued functions) such that

$$\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \boldsymbol{s} \ \text{ in } D, \qquad \mu p + \nabla\cdot\boldsymbol{\sigma} = f \ \text{ in } D. \tag{56.15}$$

Typical SI units are $\mathrm{m}\cdot\mathrm{s}^{-1}$ for $\boldsymbol{\sigma}$, Pa for p, $\mathrm{m}^2\cdot(\mathrm{Pa}\cdot\mathrm{s})^{-1}$ for $\mathbb{d}$, $\mathrm{Pa}\cdot\mathrm{m}^{-1}$ for $\boldsymbol{s}$, $(\mathrm{Pa}\cdot\mathrm{s})^{-1}$ for μ, and s^{-1} for f. It is possible to eliminate $\boldsymbol{\sigma}$ from (56.15), and one then obtains the diffusion-reaction problem $\mu p - \nabla\cdot(\mathbb{d}\nabla p) = f - \nabla\cdot(\mathbb{d}\boldsymbol{s})$ in D. But here, as in Chapter 51, we want to retain both dependent variables and work with the $\mathbb{R}^{d+1}$-valued function $u := (\boldsymbol{\sigma}, p)$. We recover Friedrichs' formalism by setting $m := d+1$ and

$$\mathcal{K} := \left[\begin{array}{c:c} \mathbb{d}^{-1} & \mathbb{O}_{d\times 1} \\ \hdashline \mathbb{O}_{1\times d} & \mu \end{array}\right], \qquad \mathcal{A}^k := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{e}_k \\ \hdashline (\boldsymbol{e}_k)^\mathsf{T} & 0 \end{array}\right], \ \forall k \in \{1{:}d\},$$

where $\boldsymbol{e}_k$ is the k-th vector of the canonical Cartesian basis of $\mathbb{R}^d$ and $\mathbb{O}_{s\times t}$ the zero matrix in $\mathbb{R}^{s\times t}$. The assumption (56.1a) is satisfied since $\mu \in L^\infty(D)$, $\lambda_\flat > 0$, and $\mathcal{X} = \mathbb{O}_{m\times m}$. The assumption (56.1b) is satisfied by construction. Finally, the assumption (56.1c) is satisfied provided we assume that

$$\mu_\flat := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \mu(\boldsymbol{x}) > 0.$$

The boundary field $\mathcal{N}$ is

$$\mathcal{N} := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n} \\ \hdashline \boldsymbol{n}^\mathsf{T} & 0 \end{array}\right].$$

The integration by parts formula (56.5) is a reformulation of the identity

$$\int_D \big(\nabla p\cdot\boldsymbol{\tau} + p(\nabla\cdot\boldsymbol{\tau})\big)\,\mathrm{d}x = \int_{\partial D} p(\boldsymbol{\tau}\cdot\boldsymbol{n})\,\mathrm{d}s.$$

The Dirichlet condition $p_{|\partial D} = 0$ and the Neumann condition $\boldsymbol{\sigma}_{|\partial D}\cdot\boldsymbol{n} = 0$ can be enforced, respectively, by using the boundary fields

$$\mathcal{M}_\mathrm{d} := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & -\boldsymbol{n} \\ \hdashline \boldsymbol{n}^\mathsf{T} & 0 \end{array}\right], \qquad \mathcal{M}_\mathrm{n} := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \boldsymbol{n} \\ \hdashline -\boldsymbol{n}^\mathsf{T} & 0 \end{array}\right].$$

Indeed, setting $(\mathcal{M}_\mathrm{d} - \mathcal{N})u = 0$ with $u := (\boldsymbol{\sigma}, p)$ amounts to $p\boldsymbol{n} = \boldsymbol{0}$, i.e., $p = 0$. Being skew-symmetric, the matrix $\mathcal{M}_\mathrm{d}$ is nonnegative, i.e., (56.7a) is satisfied. Moreover, the property (56.7b)

results from $\ker(\mathcal{M}_\mathrm{d}-\mathcal{N})=\mathbb{R}^d\times\{0\}$ and $\ker(\mathcal{M}_\mathrm{d}+\mathcal{N})\supset\{\mathbf{0}\}\times\mathbb{R}$. Similar arguments can be invoked for $\mathcal{M}_\mathrm{n}$. Finally, for all $v:=(\boldsymbol{\sigma},p)\in C^1(\overline{D};\mathbb{R}^m)$ s.t. either $(\mathcal{M}_\mathrm{d}-\mathcal{N})v_{|\partial D}=0$ or $(\mathcal{M}_\mathrm{n}-\mathcal{N})v_{|\partial D}=0$, the L-coercivity property (56.8) with $L:=L^2(D;\mathbb{R}^{d+1})$ becomes

$$(A(\boldsymbol{\sigma},p),(\boldsymbol{\sigma},p))_L\geq\lambda_\sharp^{-1}\|\boldsymbol{\sigma}\|^2_{\boldsymbol{L}^2(D)}+\mu_\flat\|p\|^2_{L^2(D)}. \tag{56.16}$$

Notice that $(\lambda_\sharp/\mu_\flat)^{\frac12}$ is a length scale, and the unit in (56.16) is $\mathrm{J\cdot s^{-1}}$ (recall that $\mathrm{Pa}=\mathrm{J\cdot m^{-3}}$).

56.2.3 Maxwell's equations

Let D be a Lipschitz domain in $\mathbb{R}^3$. We consider the time-harmonic version of Maxwell's equations in the low-frequency regime where the displacement currents are negligible; see §43.1. Let σ be the electrical conductivity, μ the magnetic permeability, $\omega>0$ the angular frequency, and $\mathsf{i}^2=-1$. We assume that $\mu,\sigma\in L^\infty(D)$, and for simplicity we assume that both μ and σ are real-valued and nonnegative. Given $\boldsymbol{j}_\mathrm{s}\in\boldsymbol{L}^2(D):=L^2(D;\mathbb{C}^3)$, we seek the fields $\boldsymbol{E}:D\to\mathbb{C}^3$ and $\boldsymbol{H}:D\to\mathbb{C}^3$ satisfying Ampère's and Faraday's laws:

$$\sigma\boldsymbol{E}-\nabla\times\boldsymbol{H}=-\boldsymbol{j}_\mathrm{s}\ \text{ in } D,\qquad \mathsf{i}\omega\mu\boldsymbol{H}+\nabla\times\boldsymbol{E}=\mathbf{0}\ \text{ in } D. \tag{56.17}$$

Typical SI units are $\mathrm{J\cdot(A\cdot s\cdot m)^{-1}}$ for $\boldsymbol{E}$, $\mathrm{A\cdot m^{-1}}$ for $\boldsymbol{H}$, $\mathrm{A^2\cdot s\cdot(J\cdot m)^{-1}}$ for σ, $\mathrm{s^{-1}}$ for ω, $\mathrm{J\cdot(m\cdot A^2)^{-1}}$ for μ, and $\mathrm{A\cdot m^{-2}}$ for $\boldsymbol{j}_\mathrm{s}$. Notice that having a nonzero right-hand side in the second equation in (56.17) would not change the structure of the problem. Contrary to Chapter 43, here we do not eliminate one of the unknown fields from (56.17), i.e., we are going to work with the $\mathbb{C}^6$-valued dependent variable $u:=(\boldsymbol{E},\boldsymbol{H})$. Let $\theta\in[0,2\pi)$ be a number (arbitrary for the time being). Let us multiply Ampère's law by $e^{\mathsf{i}\theta}$ and Faraday's law by $e^{-\mathsf{i}\theta}$. We recover Friedrichs' formalism by setting $m:=6$ and

$$\mathcal{K}:=\left[\begin{array}{c:c}e^{\mathsf{i}\theta}\sigma\mathbb{I}_3&\mathbb{O}_{3\times3}\\\hdashline\mathbb{O}_{3\times3}&\mathsf{i}e^{-\mathsf{i}\theta}\omega\mu\mathbb{I}_3\end{array}\right],\qquad \mathcal{A}^k:=\left[\begin{array}{c:c}\mathbb{O}_{3\times3}&-e^{\mathsf{i}\theta}\mathbb{J}^k\\\hdashline e^{-\mathsf{i}\theta}\mathbb{J}^k&\mathbb{O}_{3\times3}\end{array}\right],\ \forall k\in\{1{:}d\},$$

where $\mathbb{I}_3$ is the identity matrix in $\mathbb{C}^3$, $\mathbb{J}^k_{ij}:=\varepsilon_{ikj}$ for all $i,j,k\in\{1,2,3\}$, and ε_{ikj} is the Levi-Civita symbol ($\varepsilon_{ijk}:=0$ if at least two indices take the same value, $\varepsilon_{123}=\varepsilon_{231}=\varepsilon_{312}:=1$ (i.e., for even permutations), and $\varepsilon_{132}=\varepsilon_{213}=\varepsilon_{321}:=-1$ (i.e., for odd permutations)). Notice that $\mathbb{J}^k$ is skew-symmetric. The assumption (56.1a) is satisfied since $\sigma,\mu\in L^\infty(D)$ and $\mathcal{X}=\mathbb{O}_{6\times6}$. The assumption (56.1b) is satisfied since, $\mathbb{J}^k$ being skew-symmetric, we have $(-e^{\mathsf{i}\theta}\mathbb{J}^k)^\mathsf{H}=-e^{-\mathsf{i}\theta}(\mathbb{J}^k)^\mathsf{T}=e^{-\mathsf{i}\theta}\mathbb{J}^k$. Finally, recalling that we supposed that σ and μ are real-valued, the assumption (56.1c) is satisfied if we take $\theta:=\frac{\pi}{4}$ and assume that

$$\sigma_\flat:=\operatorname*{ess\,inf}_{\boldsymbol{x}\in D}\sigma(\boldsymbol{x})>0,\qquad \mu_\flat:=\operatorname*{ess\,inf}_{\boldsymbol{x}\in D}\mu(\boldsymbol{x})>0. \tag{56.18}$$

We take $\theta:=\frac{\pi}{4}$ in the rest of this section (see Example 43.2 for a more general setting). The boundary field $\mathcal{N}$ is

$$\mathcal{N}:=\left[\begin{array}{c:c}\mathbb{O}_{3\times3}&e^{\mathsf{i}\theta}\mathbb{T}\\\hdashline -e^{-\mathsf{i}\theta}\mathbb{T}&\mathbb{O}_{3\times3}\end{array}\right],$$

where $\mathbb{T}_{ij}:=\sum_{k\in\{1:3\}}n_k\varepsilon_{ijk}$ for all $i,j\in\{1,2,3\}$. Notice that the definition of $\mathbb{T}$ implies that $\mathbb{T}\boldsymbol{\xi}=\boldsymbol{\xi}\times\boldsymbol{n}$ for all $\boldsymbol{\xi}\in\mathbb{C}^3$. The integration by parts formula (56.5) is a reformulation of the identity

$$\int_D\big(\boldsymbol{b}\cdot(\nabla\times\boldsymbol{E})-\boldsymbol{E}\cdot(\nabla\times\boldsymbol{b})\big)\,\mathrm{d}x=\int_{\partial D}\boldsymbol{b}\cdot(\boldsymbol{n}\times\boldsymbol{E})\,\mathrm{d}s.$$

The boundary conditions $\boldsymbol{H}_{|\partial D}\times\boldsymbol{n}=\boldsymbol{0}$ and $\boldsymbol{E}_{|\partial D}\times\boldsymbol{n}=\boldsymbol{0}$ can be enforced, respectively, by using the boundary fields

$$\mathcal{M}_H=\left[\begin{array}{c:c}\mathbb{O}_{3\times3} & -e^{\mathrm{i}\theta}\mathbb{T}\\ \hdashline -e^{-\mathrm{i}\theta}\mathbb{T} & \mathbb{O}_{3\times3}\end{array}\right],\qquad \mathcal{M}_E=\left[\begin{array}{c:c}\mathbb{O}_{3\times3} & e^{\mathrm{i}\theta}\mathbb{T}\\ \hdashline e^{-\mathrm{i}\theta}\mathbb{T} & \mathbb{O}_{3\times3}\end{array}\right].$$

Indeed, enforcing $(\mathcal{M}_H-\mathcal{N})u=0$ with $u:=(\boldsymbol{E},\boldsymbol{H})$ amounts to $\mathbb{T}\boldsymbol{H}=\boldsymbol{0}$, i.e., $\boldsymbol{H}\times\boldsymbol{n}=\boldsymbol{0}$. The matrix $\mathcal{M}_H$ is nonnegative since it is skew-symmetric, i.e., (56.7a) is satisfied. Moreover, the property (56.7b) results from $\ker(\mathcal{M}_H-\mathcal{N})=\mathbb{C}^3\times\operatorname{span}\{\boldsymbol{n}\}$ and $\ker(\mathcal{M}_H+\mathcal{N})=\operatorname{span}\{\boldsymbol{n}\}\times\mathbb{C}^3$. Similar arguments can be invoked for $\mathcal{M}_E$. Finally, for all $v:=(\boldsymbol{E},\boldsymbol{H})\in C^1(\overline{D};\mathbb{C}^6)$ s.t. either $(\mathcal{M}_H-\mathcal{N})v_{|\partial D}=0$ or $(\mathcal{M}_E-\mathcal{N})v_{|\partial D}=0$, the L-coercivity property (56.8) with $L:=L^2(D;\mathbb{C}^6)$ becomes

$$\Re\big((A(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{E},\boldsymbol{H}))_L\big)\geq\frac{1}{\sqrt{2}}\big(\sigma_\flat\|\boldsymbol{E}\|^2_{\boldsymbol{L}^2(D)}+\omega\mu_\flat\|\boldsymbol{H}\|^2_{\boldsymbol{L}^2(D)}\big).\tag{56.19}$$

Notice that $(\sigma_\flat\omega\mu_\flat)^{-\frac{1}{2}}$ is a length scale, and the unit in (56.19) is $\mathrm{J}\cdot\mathrm{s}^{-1}$.

56.3 Weak formulation and well-posedness

The aim of this section is to devise a weak formulation of Friedrichs' systems for which the well-posedness can be established by using the BNB theorem (Theorem 25.9). The material is inspired from a series of papers by the authors [118, 119, 120].

56.3.1 Minimal domain, maximal domain, and graph space

We consider the space $\mathcal{Y}:=C_0^\infty(D;\mathbb{C}^m)$ composed of the smooth $\mathbb{C}^m$-valued fields compactly supported in D, and the Hilbert space $L:=L^2(D;\mathbb{C}^m)$, which we use as pivot space (i.e., we identify L and its dual space). Although other functional settings can be considered (see §24.2.1 for an example with Banach spaces), the prominent role played by L^2 is motivated by a large class of stabilized finite element techniques studied in the forthcoming chapters.

Let us define the inner product

$$(\cdot,\cdot)_V:=\mu_0(\cdot,\cdot)_L+\mu_0^{-1}(A_1(\cdot),A_1(\cdot))_L,\tag{56.20}$$

and let the induced norm be denoted by $\|\cdot\|_V$ (the scaling factors μ_0 and μ_0^{-1} are introduced so that the two terms composing the inner product have coherent units). Let $V_\mathcal{Y}$ be the completion of $\mathcal{Y}$ with respect to the norm $\|\cdot\|_V$, i.e., $V_\mathcal{Y}:=\overline{\mathcal{Y}}^V$. Using L as pivot space gives

$$\mathcal{Y}\subset V_\mathcal{Y}\hookrightarrow L\equiv L'\hookrightarrow V_\mathcal{Y}'\subset\mathcal{Y}',\tag{56.21}$$

where $\mathcal{Y}'$ is the algebraic dual of $\mathcal{Y}$ and L', $V_\mathcal{Y}'$ are topological dual spaces. Let us set $\tilde{A}_1(v):=-\mathcal{X}v-A_1(v)$ for all $v\in\mathcal{Y}$. A density argument shows that the operators A_1 and $\tilde{A}_1$ can be extended to bounded linear operators $A_1,\tilde{A}_1:V_\mathcal{Y}\to L$ (we use the same notation for A_1 and $\tilde{A}_1$). Following Aubin [16, §5.5], we say that $V_\mathcal{Y}$ is the *minimal domain* of A_1 and $\tilde{A}_1$ (or A and $\tilde{A}$). One integration by parts and a density argument show that $(A_1(\phi),\psi)_L=(\phi,\tilde{A}_1(\psi))_L$ for all $\phi,\psi\in V_\mathcal{Y}$. For any $v\in L$, $A_1(v)$ can be defined in $V_\mathcal{Y}'$ by setting $\langle A_1(v),\phi\rangle_{V_\mathcal{Y}',V_\mathcal{Y}}:=(v,\tilde{A}_1(\phi))_L$ for all $\phi\in V_\mathcal{Y}$. This definition extends $A_1:V_\mathcal{Y}\to L$ to a bounded linear operator $A_1:L\to V_\mathcal{Y}'$ (we

use the same notation for A_1). Since $L \hookrightarrow V_{\mathcal{Y}}'$, it makes sense to define the following space which we call *graph space*:

$$V := \{v \in L;\ A_1(v) \in L\}, \tag{56.22}$$

where $A_1(v) \in L$ means that

$$\sup_{\phi \in V_{\mathcal{Y}}} \frac{|\langle A_1(v), \phi\rangle_{V_{\mathcal{Y}}', V_{\mathcal{Y}}}|}{\|\phi\|_L} := \sup_{\phi \in V_{\mathcal{Y}}} \frac{|(v, \tilde{A}_1(\phi))_L|}{\|\phi\|_L} < \infty. \tag{56.23}$$

Similarly, we define the extension $\tilde{A}_1 : L \to V_{\mathcal{Y}}'$ by setting $\langle \tilde{A}_1(v), \phi\rangle_{V_{\mathcal{Y}}', V_{\mathcal{Y}}} := (v, A_1(\phi))_L$ for all $v \in L$ and all $\phi \in V_{\mathcal{Y}}$. Still following [16], we say that V is the *maximal domain* of A_1 and $\tilde{A}_1$ (or A and $\tilde{A}$).

Proposition 56.4 (Hilbert space). *The graph space V is a Hilbert space when equipped with the inner product $(\cdot,\cdot)_V$.*

Proof. Let $(v_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in the graph space V. Then $(v_n)_{n\in\mathbb{N}}$ and $(A_1(v_n))_{n\in\mathbb{N}}$ are Cauchy sequences in L. Denote by v and w the respective limits in L. Since $(v_n, \tilde{A}_1(\phi))_L = \langle A_1(v_n), \phi\rangle_{V_{\mathcal{Y}}', V_{\mathcal{Y}}} = (A_1(v_n), \phi)_L$ for all $\phi \in V_{\mathcal{Y}}$, we infer that

$$(v, \tilde{A}_1(\phi))_L \underset{n\to\infty}{\longleftarrow} (v_n, \tilde{A}_1(\phi))_L = (A_1(v_n), \phi)_L \underset{n\to\infty}{\longrightarrow} (w, \phi)_L.$$

Hence, $\langle A_1(v), \phi\rangle_{V_{\mathcal{Y}}', V_{\mathcal{Y}}} = (v, \tilde{A}_1(\phi))_L = (w, \phi)_L$, proving that $A_1(v)$ is in L with $A_1(v) = w$. □

In conclusion, the above argumentation has lead us to introduce $V_{\mathcal{Y}}$, which we call minimal domain of A_1 ($\tilde{A}_1$, A, or $\tilde{A}$), and V, which we call maximal domain (or graph space) of A_1. We have shown that V is a Hilbert space. We have extended the operators A_1, $\tilde{A}_1$, A, and $\tilde{A}$, initially defined on $\mathcal{Y}$ only, to bounded operators from V to L:

$$A_1 \in \mathcal{L}(V; L), \quad \tilde{A}_1 \in \mathcal{L}(V; L), \quad A \in \mathcal{L}(V; L), \quad \tilde{A} \in \mathcal{L}(V; L). \tag{56.24}$$

Example 56.5 (Transport). Let $D := (0,1)^2$ and $A_1(v) := e^x \partial_x v$, so that $\tilde{A}_1(v) = -e^x v - e^x \partial_x v$. Then the minimal domain is $V_{\mathcal{Y}} := \{v \in L^2(D) \mid \partial_x v \in L^2(D),\ v(0,y) = 0,\ v(1,y) = 0, \forall y \in (0,1)\}$, and the maximal domain is $V := \{v \in L^2(D) \mid \partial_x v \in L^2(D)\}$. When solving the first-order PDE $v + A_1(v) = f$, one enforces an homogeneous Dirichlet condition on the inflow boundary $\{x = 0\}$, i.e., one seeks the solution in $V_0 := \{v \in V \mid v(0,y) = 0, \forall y \in (0,1)\}$. Notice that $V_{\mathcal{Y}} \subsetneq V_0 \subsetneq V$. □

Remark 56.6 (Density). It is shown in Jensen [197, p. 21] that the space $C^\infty(\overline{D}; \mathbb{C}^m)$ is dense in the maximal domain V of the operators (recall that $C_0^\infty(D; \mathbb{C}^m)$ is by definition dense in the minimal domain $V_{\mathcal{Y}}$). □

56.3.2 The boundary operators N and M

Since A_1 is a first-order differential operator, defining the trace at the boundary of a function in the graph space V is not straightforward. For any $v \in V$, the trace $\mathcal{N} v_{|\partial D}$ can be given a meaning in $H^{-\frac{1}{2}}(\partial D; \mathbb{C}^m)$; see Rauch [241], Jensen [197]. Recall that $\gamma^{\mathrm{g}} : H^1(D; \mathbb{C}^m) \to H^{\frac{1}{2}}(\partial D; \mathbb{C}^m)$ is surjective (here, γ^{g} is the $\mathbb{C}^m$-valued version of the scalar trace operator introduced in Theorem 3.10), and let $(\gamma^{\mathrm{g}})^\dagger$ be any right inverse of γ^{g}. Then the action of $\mathcal{N} v_{|\partial D}$ on $H^{\frac{1}{2}}(\partial D; \mathbb{C}^m)$ can de defined by setting

$$\langle \mathcal{N} v, s\rangle_{H^{-\frac{1}{2}}(\partial D; \mathbb{C}^m), H^{\frac{1}{2}}(\partial D; \mathbb{C}^m)} := (A(v), (\gamma^{\mathrm{g}})^\dagger(s))_L - (v, \tilde{A}((\gamma^{\mathrm{g}})^\dagger(s)))_L,$$

for all $s \in H^{\frac{1}{2}}(\partial D; \mathbb{C}^m)$. This construction is explained with more details in §4.3 for $A_1(\boldsymbol{v}) := \nabla \times \boldsymbol{v}$ and $A_1(v) := \nabla \cdot \boldsymbol{v}$. This meaning is however not suitable for the weak formulation we have in mind. This is why we now introduce two additional operators N and M to replace the boundary fields $\mathcal{N}$ and $\mathcal{M}$. We define the operator $N \in \mathcal{L}(V; V')$ by (compare with (56.5))

$$\langle N(v), w\rangle_{V',V} := (A(v), w)_L - (v, \tilde{A}(w))_L, \qquad \forall v, w \in V. \tag{56.25}$$

This definition makes sense since both A and $\tilde{A}$ are in $\mathcal{L}(V; L)$. Moreover, the operator N is self-adjoint since (56.25) can be rewritten as

$$\langle N(v), w\rangle_{V',V} = (\mathcal{X} v, w)_L + (A_1(v), w)_L + (v, A_1(w))_L, \tag{56.26}$$

so that $\langle N(v), w\rangle_{V',V} = \overline{\langle N(w), v\rangle_{V',V}}$. Furthermore, we have $V_{\mathcal{Y}} \subset \ker(N)$ and $\mathrm{im}(N) \subset V_{\mathcal{Y}}^{\perp} = \{v' \in V' \mid \forall \phi \in V_{\mathcal{Y}}, \langle v', \phi\rangle_{V'_{\mathcal{Y}}, V_{\mathcal{Y}}} = 0\}$. Actually, as shown in [123], the following holds true:

$$\ker(N) = V_{\mathcal{Y}}, \qquad \mathrm{im}(N) = V_{\mathcal{Y}}^{\perp}.$$

The fact that $\ker(N) = V_{\mathcal{Y}}$ means that N is a *boundary operator*.

Boundary conditions in Friedrichs' systems can be formulated by assuming that there exists an operator $M \in \mathcal{L}(V; V')$ such that

$$M \text{ is monotone, i.e., } |v|_M^2 := \Re\big(\langle M(v), v\rangle_{V',V}\big) \geq 0 \text{ for all } v \in V, \tag{56.27a}$$
$$\ker(N - M) + \ker(N + M) = V. \tag{56.27b}$$

Let $M^* \in \mathcal{L}(V; V')$ denote the adjoint operator of M, i.e., $\langle M^*(w), v\rangle_{V',V} = \overline{\langle M(v), w\rangle_{V',V}}$. It is proved in [123] that, under the assumptions (56.27),

$$\ker(N) = \ker(M) = \ker(M^*),$$
$$\mathrm{im}(N) = \mathrm{im}(M) = \mathrm{im}(M^*).$$

In particular, M is a *boundary operator* just like N.

Remark 56.7 (Other formalisms). A different viewpoint based on Lax's idea consisting of enforcing maximal boundary conditions by a cone technique is explored in [123]. The equivalence between this formalism and the M-based formalism (56.27) and relations with the approach based on boundary fields can be found in Antonić and Burazin [11, 12, 13]. □

56.3.3 Well-posedness

Let us set $V_0 := \ker(M - N)$ so that $V_{\mathcal{Y}} \subseteq V_0 \subseteq V$. Given $f \in L$, the problem we want to solve (compare with (56.9)) consists of seeking

$$u \in V_0 = \ker(M - N) \text{ such that } A(u) = f \text{ in } L. \tag{56.28}$$

To recast this problem into a weak form, we introduce the sesquilinear form

$$a(v, w) := (A(v), w)_L, \qquad \forall (v, w) \in V \times L.$$

Letting $\ell(w) := (f, w)_L$, we consider the following weak problem:

$$\begin{cases} \text{Find } u \in V_0 \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in L. \end{cases} \tag{56.29}$$

Lemma 56.8 (L-coercivity). *Assume* (56.1). *Let N be defined in* (56.25). *The following holds true:*

$$\Re\big(a(v,v)\big) \geq \mu_0\|v\|_L^2 + \frac{1}{2}\langle Nv, v\rangle_{V',V}, \qquad \forall v \in V. \tag{56.30}$$

Moreover, let M satisfy (56.27). *The following holds true:*

$$\Re\big(a(v,v)\big) \geq \mu_0\|v\|_L^2 + \frac{1}{2}|v|_M^2 \geq \mu_0\|v\|_L^2, \qquad \forall v \in V_0. \tag{56.31}$$

Proof. One proceeds as in the proof of Lemma 56.2 to establish (56.30). Moreover, (56.31) follows from (56.30), the definition $V_0 := \ker(M-N)$, and the monotonicity of M. □

Theorem 56.9 (Well-posedness). *Assume* (56.1) *and* (56.27). *Then the model problem* (56.29) *is well-posed, i.e., $A : V_0 \to L$ is an isomorphism.*

Proof. Since V_0 and L are Hilbert spaces and since a and ℓ are bounded on $V_0\times L$ and L, respectively, we just have to verify that the two conditions of the BNB theorem are satisfied.
(1) Proof of (BNB1). Let us set $S(v) := \sup_{w\in L}\frac{|a(v,w)|}{\|w\|_L}$ for all $v \in V_0$. We want to prove that there exists $\alpha > 0$ such that

$$\alpha\|v\|_V \leq S(v), \qquad \forall v \in V_0. \tag{56.32}$$

Let $v \in V_0$. Using (56.31), we infer that $\|v\|_L \leq \frac{1}{\mu_0}\frac{\Re(a(v,v))}{\|v\|_L} \leq \frac{1}{\mu_0}S(v)$. Using the triangle inequality and letting $\mu_\infty := \|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}$, we obtain

$$\|A_1(v)\|_L \leq \|\mathcal{K}v\|_L + \|A(v)\|_L \leq \mu_\infty\|v\|_L + \|A(v)\|_L \leq \Big(\frac{\mu_\infty}{\mu_0}+1\Big)S(v),$$

thus yielding (56.32) with $\alpha := \mu_0^{\frac{1}{2}}\big(1+(1+\frac{\mu_\infty}{\mu_0})^2\big)^{-\frac{1}{2}}$.
(2) Proof of (BNB2). Let $w \in L$ be such that $a(v,w) = 0$ for all $v \in V_0$, and let us prove that $w = 0$. Recalling that $V_{\mathcal{Y}} = \ker(N) = \ker(M) \subset V_0$ and the definition of $\tilde{A}$, we infer that the following holds true for all $\phi \in V_{\mathcal{Y}} \subset V_0$:

$$\overline{\langle\tilde{A}(w),\phi\rangle_{V_{\mathcal{Y}}',V_{\mathcal{Y}}}} = (A(\phi),w)_L = a(\phi,w) = 0.$$

Hence, $\tilde{A}(w) = 0$, thereby showing that $w \in V$. Owing to the properties satisfied by w, we also infer that

$$\langle N(v),w\rangle_{V',V} = (A(v),w)_L - (v,\tilde{A}(w))_L = 0 - 0 = 0, \tag{56.33}$$

for all $v \in V_0$. Let us now show that $w \in \ker(M^*+N)$. For all $v \in V$, using (56.27b) to write $v = v_+ + v_-$ with $v_\pm \in \ker(M\pm N)$, we obtain

$$\begin{aligned}\overline{\langle(M^*+N)(w),v\rangle_{V',V}} &= \langle(M+N)(v),w\rangle_{V',V} = \langle(M+N)(v_+ + v_-),w\rangle_{V',V}\\ &= \langle(M+N)(v_-),w\rangle_{V',V} = 2\langle N(v_-),w\rangle_{V',V} = 0,\end{aligned}$$

owing to the self-adjointness of N, the identity $N(v_-) = M(v_-)$ since $v_- \in \ker(M-N)$, and the property (56.33) since $v_- \in V_0$. Since v is arbitrary in V, this shows that $w \in \ker(M^*+N)$. Using (56.25) and (56.30) together with $N(w) = -M^*(w)$, we obtain

$$\begin{aligned}\Re\big((w,\tilde{A}(w))_L\big) &= \Re\big((A(w),w)_L\big) - \langle N(w),w\rangle_{V',V}\\ &\geq \mu_0\|w\|_L^2 - \frac{1}{2}\Re\big(\langle N(w),w\rangle_{V',V}\big) = \mu_0\|w\|_L^2 + \frac{1}{2}|w|_M^2 \geq \mu_0\|w\|_L^2.\end{aligned}$$

Recalling that $\tilde{A}(w) = 0$, the above inequality implies that $w = 0$. □

Remark 56.10 (Graph-norm estimate). In the proof of Theorem 56.9, we used that $\|A(v)\|_L = \sup_{w\in L} \frac{|a(v,w)|}{\|w\|_L}$. Since the supremum is attained for $w := A(v)$, this identity shows that the control on the graph norm of v follows from the fact that we can use $A_1(v) = A(v) - \mathcal{K}v$ as a test function. The whole difficulty of approximating first-order PDEs (see Chapters 57 to 61) has its roots in this observation. □

Remark 56.11 (Partial positivity). The positivity assumption (56.1c) can be relaxed if the missing control on $\|v\|_L$ can be recovered from an estimate on $\|A_1(v)\|_L$. This is possible in the context of elliptic PDEs in mixed form by invoking a Poincaré-type inequality; see Exercise 56.5. □

Remark 56.12 (Localization). Let us define the operator $K \in \mathcal{L}(L;L)$ s.t. $K(v) := \mathcal{K}v$, which means that $K(v)(\boldsymbol{x}) = \mathcal{K}(v(\boldsymbol{x}))$ for all $\boldsymbol{x} \in D$. One says that K is a local operator. Everything that is said in this chapter and the following chapters holds true for nonlocal operators as well. More precisely, we can assume that $A = K + A_1$, where K is any bounded operator on L satisfying the assumption $((K+K^*)(v) - \mathcal{X}v, v)_L \geq 2\mu_0\|v\|_L^2$. The formal adjoint $\tilde{A}$ is then defined by $\tilde{A}(v) := K^*(v) - \mathcal{X}v + A_1(v)$. Such nonlocal operators are found in the Boltzmann equation and in the neutron transport equation. In this context, K is usually called collision operator; see Exercise 56.3 for an application to the neutron transport equation. □

56.3.4 Examples

Example 56.13 (Advection-reaction). The bilinear form a associated with the model advection-reaction equation is defined by setting

$$a(v,w) := \int_D (\mu vw + (\boldsymbol{\beta}\cdot\nabla v)w)\,\mathrm{d}x,$$

for all $v \in V$ and all $w \in L^2(D;\mathbb{R})$, with $V := \{v \in L^2(D;\mathbb{R}) \mid \boldsymbol{\beta}\cdot\nabla v \in L^2(D;\mathbb{R})\}$. Moreover, $\langle N(v), w\rangle_{V',V} = \int_D \nabla\cdot(\boldsymbol{\beta}vw)\,\mathrm{d}x$. A result on the traces of functions in V is needed to link the boundary operator N with the boundary field $\mathcal{N} := \boldsymbol{\beta}\cdot\boldsymbol{n}$. Such a result is not straightforward since the trace theorem (Theorem 3.10) for functions in $H^1(D)$ cannot be applied. It is shown in [118] that if the inflow and outflow boundaries are well-separated, i.e.,

$$\min_{(\boldsymbol{x},\boldsymbol{y})\in\partial D^-\times\partial D^+} \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell^2(\mathbb{R}^d)} > 0, \tag{56.34}$$

then the trace operator $\gamma : C^0(\overline{D}) \to C^0(\partial D)$ s.t. $\gamma(v) = v_{|\partial D}$ can be extended to a bounded linear operator from V to $L^2_{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}(\partial D;\mathbb{R})$, where the subscript $|\boldsymbol{\beta}\cdot\boldsymbol{n}|$ means that the measure $\mathrm{d}s$ is replaced by $|\boldsymbol{\beta}\cdot\boldsymbol{n}|\,\mathrm{d}s$. This result implies that $\langle N(v), w\rangle_{V',V} = \int_{\partial D}(\mathcal{N}v)w\,\mathrm{d}s$ for all $v, w \in V$. Furthermore the inflow boundary condition (56.13) can be enforced by means of the boundary operator $M \in \mathcal{L}(V;V')$ defined by $\langle M(v), w\rangle_{V',V} := \int_{\partial D}(\mathcal{M}v)w\,\mathrm{d}s$ with $\mathcal{M} := |\boldsymbol{\beta}\cdot\boldsymbol{n}|$. This operator satisfies (56.27), and we have $|v|_M = (\int_{\partial D}|\boldsymbol{\beta}\cdot\boldsymbol{n}|v^2\,\mathrm{d}s)^{\frac{1}{2}}$. One can also construct M without invoking the boundary field $\mathcal{M}$. Since the inflow and outflow boundaries are well-separated, there exists $\alpha \in C^\infty(\mathbb{R}^d)$ s.t. $\alpha_{|\partial D^-} = 0$ and $\alpha_{|\partial D^+} = 1$. Then one can set $\langle M(v), w\rangle_{V',V} := \int_D \nabla\cdot(\alpha\boldsymbol{\beta}vw)\,\mathrm{d}x$. Notice also that the separation assumption (56.34) cannot be circumvented if one wishes to work with traces in $L^2_{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}(\partial D;\mathbb{R})$; see Exercise 56.8. Alternatively, as shown in Joly [203, Thm. 2], traces and the above integration by parts formula can be defined by using principal values. □

Example 56.14 (Darcy). The bilinear form a associated with Darcy's equations is defined by setting

$$a(v,w) := \int_D \big((\mathbb{d}^{-1}\boldsymbol{\sigma})\cdot\boldsymbol{\tau} + \mu pq + \nabla p\cdot\boldsymbol{\tau} + (\nabla\cdot\boldsymbol{\sigma})q\big)\,\mathrm{d}x,$$

for all $v := (\boldsymbol{\sigma}, p) \in V$ and all $w := (\boldsymbol{\tau}, q) \in L$, with $V := \boldsymbol{H}(\mathrm{div}; D)\times H^1(D)$. This functional setting corresponds to that of §24.1.2 rather than that of §51.1. The definition of N gives $\langle N(v), w\rangle_{V',V} = \int_D \nabla\cdot(p\boldsymbol{\tau} + q\boldsymbol{\sigma})\,\mathrm{d}x$. Since fields in $\boldsymbol{H}(\mathrm{div}; D)$ have a normal trace in $H^{-\frac12}(\partial D)$ owing to Theorem 4.15, and functions in $H^1(D)$ have a trace in $H^{\frac12}(\partial D)$, letting $\langle\cdot,\cdot\rangle_{H^{-\frac12},H^{\frac12}}$ denote the duality pairing between $H^{-\frac12}(\partial D)$ and $H^{\frac12}(\partial D)$, the boundary operator N has also the following representation:

$$\langle N(\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q)\rangle_{V',V} := \langle\boldsymbol{\sigma}\cdot\boldsymbol{n}, q\rangle_{H^{-\frac12},H^{\frac12}} + \langle\boldsymbol{\tau}\cdot\boldsymbol{n}, p\rangle_{H^{-\frac12},H^{\frac12}}.$$

(One should write $\langle N(\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q)\rangle_{V',V} := \langle\gamma^{\mathrm{d}}(\boldsymbol{\sigma}), \gamma^{\mathrm{g}}(q)\rangle_{H^{-\frac12},H^{\frac12}} + \langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), \gamma^{\mathrm{g}}(p)\rangle_{H^{-\frac12},H^{\frac12}}$.) The Dirichlet boundary condition $p_{|\partial D} = 0$ can be enforced by means of the boundary operator

$$\langle M(\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q)\rangle_{V',V} := \langle\boldsymbol{\sigma}\cdot\boldsymbol{n}, q\rangle_{H^{-\frac12},H^{\frac12}} - \langle\boldsymbol{\tau}\cdot\boldsymbol{n}, p\rangle_{H^{-\frac12},H^{\frac12}}. \tag{56.35}$$

This operator satisfies (56.27); see Exercise 56.7(i). Notice that $|(\boldsymbol{\sigma}, p)|_M = 0$ for all $(\boldsymbol{\sigma}, p) \in V$. We also have $\langle M(v), w\rangle_{V',V} = \int_D \nabla\cdot(q\boldsymbol{\sigma} - p\boldsymbol{\tau})\,\mathrm{d}x$. □

Example 56.15 (Maxwell). For Maxwell's equations, the sesquilinear form a is defined by setting

$$a(v,w) := \int_D \big(e^{\mathrm{i}\theta}\sigma\boldsymbol{E}\cdot\overline{\boldsymbol{e}} + \mathrm{i}e^{-\mathrm{i}\theta}\omega\mu\boldsymbol{H}\cdot\overline{\boldsymbol{b}} - e^{\mathrm{i}\theta}(\nabla\times\boldsymbol{H})\cdot\overline{\boldsymbol{e}} + e^{-\mathrm{i}\theta}(\nabla\times\boldsymbol{E})\cdot\overline{\boldsymbol{b}}\big)\,\mathrm{d}x,$$

for all $v := (\boldsymbol{E}, \boldsymbol{H}) \in V$ and all $w := (\boldsymbol{e}, \boldsymbol{b}) \in L$ (notice that we use the Euclidean dot product and write the complex conjugate explicitly), with $V := \boldsymbol{H}(\mathrm{curl}; D)\times\boldsymbol{H}(\mathrm{curl}; D)$. Recalling the identity $\nabla\cdot(\boldsymbol{A}\times\overline{\boldsymbol{a}}) = (\nabla\times\boldsymbol{A})\cdot\overline{\boldsymbol{a}} - \boldsymbol{A}\cdot(\nabla\times\overline{\boldsymbol{a}})$, the definition of the boundary operator N gives

$$\langle N(\boldsymbol{E}, \boldsymbol{H}), (\boldsymbol{e}, \boldsymbol{b})\rangle_{V',V} := \int_D \nabla\cdot(e^{-\mathrm{i}\theta}\boldsymbol{E}\times\overline{\boldsymbol{b}} - e^{\mathrm{i}\theta}\boldsymbol{H}\times\overline{\boldsymbol{e}})\,\mathrm{d}x.$$

Owing to Theorem 4.15 (with $p := 2$), fields in $\boldsymbol{H}(\mathrm{curl}; D)$ have a tangential trace in $\boldsymbol{H}^{-\frac12}(\partial D)$. Hence, if $\boldsymbol{e}$ and $\boldsymbol{b}$ are both in $\boldsymbol{H}^1(D)$, we also have the following representation:

$$\langle N(\boldsymbol{E}, \boldsymbol{H}), (\boldsymbol{e}, \boldsymbol{b})\rangle_{V',V} := e^{\mathrm{i}\theta}\langle\boldsymbol{H}\times\boldsymbol{n}, \boldsymbol{e}\rangle_{\boldsymbol{H}^{-\frac12},\boldsymbol{H}^{\frac12}} - e^{-\mathrm{i}\theta}\langle\boldsymbol{E}\times\boldsymbol{n}, \boldsymbol{b}\rangle_{\boldsymbol{H}^{-\frac12},\boldsymbol{H}^{\frac12}}.$$

(One should write $\langle N(\boldsymbol{E}, \boldsymbol{H}), (\boldsymbol{e}, \boldsymbol{b})\rangle_{V',V} := e^{\mathrm{i}\theta}\langle\gamma^{\mathrm{c}}(\boldsymbol{H}), \gamma^{\mathrm{g}}(\boldsymbol{e})\rangle_{\boldsymbol{H}^{-\frac12},\boldsymbol{H}^{\frac12}} - e^{-\mathrm{i}\theta}\langle\gamma^{\mathrm{c}}(\boldsymbol{E}), \gamma^{\mathrm{g}}(\boldsymbol{b})\rangle_{\boldsymbol{H}^{-\frac12},\boldsymbol{H}^{\frac12}}$.) The boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ can be enforced by means of the boundary operator

$$\langle M(\boldsymbol{E}, \boldsymbol{H}), (\boldsymbol{e}, \boldsymbol{b})\rangle_{V',V} := \int_D \nabla\cdot(e^{-\mathrm{i}\theta}\boldsymbol{E}\times\overline{\boldsymbol{b}} + e^{\mathrm{i}\theta}\boldsymbol{H}\times\overline{\boldsymbol{e}})\,\mathrm{d}x. \tag{56.36}$$

This operator satisfies (56.27); see Exercise 56.7(ii). Notice that $|(\boldsymbol{e}, \boldsymbol{h})|_M = 0$ for all $(\boldsymbol{e}, \boldsymbol{h}) \in V$. □

Exercises

Exercise 56.1 (Robin condition). Show how to enforce the Robin boundary condition $\gamma u - \boldsymbol{\sigma}\cdot\boldsymbol{n} = 0$ on ∂D (with $\gamma \in L^\infty(\partial D)$ and $\gamma \geq 0$ a.e. on ∂D) in the framework of §56.2.2.

Exercise 56.2 (Linear elasticity). Consider the linear elasticity model from §42.1. Verify that $\mathbb{s} - \frac{1}{d+\theta}\operatorname{tr}(\mathbb{s})\mathbb{I}_d = \mu(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\mathsf{T})$ with $\theta := \frac{2\mu}{\lambda}$ and that $\frac{1}{2}\nabla\cdot(\mathbb{s} + \mathbb{s}^\mathsf{T}) + \boldsymbol{f} = \boldsymbol{0}$. Write this system using Friedrichs' formalism. (*Hint*: identify $\mathbb{s} \in \mathbb{R}^{d\times d}$ with a vector $\boldsymbol{s} \in \mathbb{R}^{d^2}$ by setting $\boldsymbol{s}_{[ij]} := \mathbb{s}_{ij}$ with $[ij] := d(j-1)+i$ for all $i,j \in \{1{:}d\}$.) Verify (56.1a)-(56.1b) and that the upper left block of $\mathcal{K}$, say $\mathcal{K}^{ss}$, is positive definite. What happens in the incompressible limit $\lambda \to \infty$?

Exercise 56.3 (Positivity, locality). (i) Reprove Theorem 56.9 by replacing the assumption made on $\mathcal{K}$ by those stated in Remark 56.12. (ii) Let $D := (0,a)\times(-1,1)$, $a > 0$, and let $K : L^2(D) \to L^2(D)$ be such that $K(v)(x,y) := v(x,y) - \frac{\sigma}{2}\int_{-1}^{+1} v(x,\xi)\,\mathrm{d}\xi$ with $\sigma \in [0,1)$. Assuming $\mathcal{X} := 0$, prove that K satisfies the assumptions from Remark 56.12.

Exercise 56.4 (Wave equation). Consider the wave equation $\frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x^2} = f$ in $D := (0,1)\times(-1,1)$ with the boundary conditions $\frac{\partial v}{\partial t}(t,\pm 1) = 0$ for all $t \in (0,1)$ and $\frac{\partial v}{\partial t}(0,x) = \frac{\partial v}{\partial x}(0,x) = 0$ for all $x \in (-1,1)$. Recast this problem as a Friedrichs' system and identify the boundary fields $\mathcal{N}$ and $\mathcal{M}$. (*Hint*: set $u := e^{-\lambda t}(\frac{\partial v}{\partial t}, \frac{\partial v}{\partial x})$ with $\lambda > 0$.)

Exercise 56.5 (Partial positivity). Assume that there is an orthogonal projection operator $\mathcal{P} \in \mathbb{C}^{m\times m}$ (i.e., $\mathcal{P}^\mathsf{T} = \mathcal{P}$ and $\mathcal{P}^2 = \mathcal{P}$) such that

$$\mathcal{K} + \mathcal{K}^\mathsf{T} - \mathcal{X} \geq 2\mu_0 \mathcal{P} \text{ a.e. in } D, \tag{56.37a}$$

$$\sup_{w\in L} \frac{|(A(v),w)_L|}{\|w\|_L} \geq \alpha\|(\mathbb{I}_m - \mathcal{P})(v)\|_L - \lambda\|\mathcal{P}(v)\|_L \text{ for all } v \in V_0, \tag{56.37b}$$

$$\|\mathcal{P}(w)\|_L \geq \gamma\|(\mathbb{I}_m - \mathcal{P})(w)\|_L \text{ for all } w \in \tilde{V}_0 \text{ s.t. } \tilde{A}(w) = 0, \tag{56.37c}$$

with $\mu_0 > 0$ $\alpha > 0$, $\gamma > 0$, λ, and $\tilde{V}_0 := \ker(M^* + N)$. (i) Assume (56.1a), (56.1b), (56.27), and (56.37). Prove that $A : V_0 \to L$ is an isomorphism. (*Hint*: adapt the proof of Theorem 56.9.) (ii) Verify (56.37a) for Darcy's equations with $\mu := 0$ and a Dirichlet boundary condition on p. (*Hint*: use a Poincaré–Steklov inequality.)

Exercise 56.6 ((BNB1) for Darcy and Maxwell). (i) Prove the condition (BNB1) for Darcy's equations with Dirichlet or Neumann condition. (*Hint*: use the test function $w := (\boldsymbol{\tau}, q) := (\boldsymbol{\sigma} + \mathrm{d}\nabla p, p + \mu^{-1}\nabla\cdot\boldsymbol{\sigma})$.) (ii) Do the same for Maxwell's equations with the condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$ or $\boldsymbol{E}\times\boldsymbol{n} = \boldsymbol{0}$. (*Hint*: use the test function $w := (\boldsymbol{e},\boldsymbol{b}) := (e^{-\mathrm{i}\theta}(\boldsymbol{E} - \mathrm{i}\frac{1}{\sigma}\nabla\times\boldsymbol{H}), e^{\mathrm{i}\theta}(\boldsymbol{H} + \frac{1}{\omega\mu}\nabla\times\boldsymbol{E}))$ where $\theta := \frac{\pi}{4}$.)

Exercise 56.7 (Boundary operator for Darcy and Maxwell). (i) Verify that M defined in (56.35) satisfies (56.27) and that it can be used to enforce a Dirichlet boundary condition on p. (*Hint*: use Theorem 4.15.) How should M be modified to enforce a Neumann condition? (ii) Verify that M defined in (56.36) satisfies (56.27) and that it can be used to enforce the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$. (*Hint*: use the surjectivity of traces from $H^1(D)$ onto $H^{\frac{1}{2}}(\partial D)$ and (4.11).) How should M be modified to enforce the boundary condition $\boldsymbol{E}\times\boldsymbol{n} = \boldsymbol{0}$?

Exercise 56.8 (Separation assumption). Let $D := \{(x_1,x_2) \in \mathbb{R}^2 \mid 0 < x_2 < 1 \text{ and } |x_1| < x_2\}$ with $\boldsymbol{\beta} := (1,0)^\mathsf{T}$. Let $V := \{v \in L^2(D) \mid \boldsymbol{\beta}\cdot\nabla v \in L^2(D)\}$. Verify that the function $u(x_1,x_2) := x_2^\alpha$ is in V for $\alpha > -1$, but $u_{|\partial D} \in L^2(|\boldsymbol{\beta}\cdot\boldsymbol{n}|;\partial D)$ only if $\alpha > -\frac{1}{2}$.

Exercise 56.9 (Semi-norm $|\cdot|_M$). Let V be a complex Hilbert space, $N, M \in \mathcal{L}(V;V')$, and let $V_0 := \ker(M - N)$. Assume $N = N^*$ and $\Re(\langle M(v),v\rangle_{V',V}) \geq 0$ for all $v \in V$. Let $|v|_M^2 := \Re(\langle M(v),v\rangle_{V',V})$ for all $v \in V$. Prove that $|\langle N(v),w\rangle_{V',V}| \leq |v|_M|w|_M$ for all $v,w \in V_0$.

Chapter 57

Residual-based stabilization

This chapter is concerned with the approximation of Friedrichs' systems using H^1-conforming finite elements. The main issue one faces in this context is to achieve stability (see (27.11) for a simple one-dimensional counterexample). As mentioned in Remark 56.10, one has to use the first derivative of the solution as a test function to control the graph norm. This possibility is lost when working with H^1-conforming finite elements, since the first derivative of the discrete solution can no longer be represented by discrete test functions. As a result, one needs to devise suitable stabilization mechanisms. Those presented in this chapter are inspired by the least-squares (LS), or minimal residual, technique from linear algebra. The LS approximation gives optimal error estimates in the graph norm, but unfortunately it gives suboptimal L^2-error estimates in most situations. The *Galerkin/least-squares* (GaLS) method improves the situation by combining the standard Galerkin approach with the LS technique and mesh-dependent weights. The GaLS method gives quasi-optimal L^2-error estimates and optimal mesh-dependent graph-norm estimates. We also show that the GaLS method can be combined with a boundary penalty technique to enforce boundary conditions weakly.

57.1 Model problem

Let us briefly recall the model problem from §56.3. We consider a Friedrichs' operator $A(v) := \mathcal{K}v + A_1(v)$, where $\mathcal{K}$ is the zeroth-order part of the operator and A_1 is the first-order part with $A_1(v) := \sum_{k\in\{1:d\}} \mathcal{A}^k \partial_k v$. We assume that the $\mathbb{C}^{m\times m}$-valued fields $\mathcal{K}$ and $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ satisfy (56.1). We have $A \in \mathcal{L}(V;L)$ with $L := L^2(D;\mathbb{C}^m)$ and the graph space $V := \{v \in L \mid A_1(v) \in L\}$ is equipped with the graph norm $\|v\|_V^2 := \mu_0\|v\|_L^2 + \mu_0^{-1}\|A_1(v)\|_L^2$, where $\mu_0 > 0$ comes from Assumption (56.1c). Let N be the boundary operator defined in (56.25) and let M be a boundary operator satisfying (56.27). Given $f \in L$ and upon setting $V_0 := \ker(M-N) \subset V$, the model problem is

$$\left\{\begin{array}{l} \text{Find } u \in V_0 \text{ such that} \\ (A(u),w)_L = (f,w)_L, \quad \forall w \in L. \end{array}\right. \tag{57.1}$$

This problem is well-posed (see Theorem 56.9). In particular, since $A : V_0 \to L$ is a bounded isomorphism, there are real numbers $0 < \alpha \le \|A\| < \infty$ s.t.

$$\alpha\,\|v\|_V \le \|A(v)\|_L \le \|A\|\,\|v\|_V, \qquad \forall v \in V_0. \tag{57.2}$$

57.2 Least-squares (LS) approximation

57.2.1 Weak problem

The LS version of problem (57.1) is as follows:

$$\begin{cases} \text{Find } u \in V_0 \text{ such that} \\ a^{\mathrm{LS}}(u,w) := (A(u),A(w))_L = (f,A(w))_L, \quad \forall w \in V_0. \end{cases} \tag{57.3}$$

Observe that the trial space and the test space are identical. Since $A : V_0 \to L$ is an isomorphism, requesting that $(A(u),A(w))_L = (f,A(w))_L$ for all $w \in V_0$ is equivalent to requesting that $(A(u),w)_L = (f,w)_L$ for all $w \in L$. Hence, the problems (57.1) and (57.3) are equivalent. The advantage of (57.3) is that its well-posedness follows from the Lax–Milgram lemma.

Proposition 57.1 (V_0-coercivity). *a^{LS} is bounded and coercive on $V_0 \times V_0$.*

Proof. The boundedness of a^{LS} follows from $|a^{\mathrm{LS}}(v,w)| \le \|A(v)\|_L \|A(w)\|_L \le \|A\|^2 \|v\|_V \|w\|_V$ for all $(v,w) \in V_0 \times V_0$, and the coercivity of a^{LS} follows from $a^{\mathrm{LS}}(v,v) = \|A(v)\|_L^2 \ge \alpha^2 \|v\|_V^2$ for all $v \in V_0$. □

Remark 57.2 (Minimal residual). Consider the functional $\mathfrak{J} : V_0 \to \mathbb{R}$ defined by $\mathfrak{J}(v) := \frac{1}{2}\|A(v) - f\|_L^2$ for all $v \in V_0$. The Fréchet derivative of $\mathfrak{J}$ is such that $D\mathfrak{J}(v)(w) = \Re\big((A(v) - f, A(w))_L\big)$ for all $w \in V_0$, i.e., the problem (57.3) is equivalent to $D\mathfrak{J}(v) = 0$ in $(V_0)'$. Since the functional $\mathfrak{J}$ is strongly convex and continuous owing to (57.2), its unique global minimizer over V_0 is the solution u to (57.3). The LS technique is well known in the context of linear algebra, where it can be traced back to Gauss (*Theoria Motus Corporum Coelestium*, 1809). More precisely, consider a linear system $\mathcal{A}U = \mathsf{B}$, where the matrix $\mathcal{A} \in \mathbb{C}^{I \times I}$ is invertible and $\mathsf{B} \in \mathbb{C}^I$ is some given vector. Left-multiplying the system by $\mathcal{A}^{\mathsf{H}}$ gives the normal equation $(\mathcal{A}^{\mathsf{H}}\mathcal{A})U = \mathcal{A}^{\mathsf{H}}\mathsf{B}$, where the matrix $\mathcal{A}^{\mathsf{H}}\mathcal{A}$ is Hermitian positive definite. □

57.2.2 Finite element setting

Our goal is to use H^1-conforming finite elements to approximate the model problem (57.3). Let $(V_h)_{h\in\mathcal{H}}$ be a sequence of H^1-conforming finite element subspaces based on a shape-regular mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly. We assume that V_h is built by using a reference finite element of degree $k \ge 1$. If $m \ge 2$, we assume for simplicity that the same reference element is used for all the components of the solution. Notice that $V_h \subset H^1(D;\mathbb{C}^m) \subset V$.

In this section (and the next one), we assume that it is possible to prescribe the boundary conditions strongly. In other words, we assume that we have at hand a subspace $V_{h0} \subset V_0$ and a quasi-interpolation operator $\mathcal{I}_{h0} : V_0 \to V_{h0}$ with optimal local approximation properties, i.e., there is c s.t.

$$\|v - \mathcal{I}_{h0}(v)\|_{L(K)} + h_K \|\nabla(v - \mathcal{I}_{h0}(v))\|_{\boldsymbol{L}(K)} \le c\, h_K^{1+r} |v|_{H^{1+r}(D_K,\mathbb{C}^m)}, \tag{57.4}$$

for all $r \in [0,k]$, all $v \in H^{1+r}(D;\mathbb{C}^m) \cap V_0$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$, with $L(K) := L^2(K;\mathbb{C}^m)$ and where $D_K := \mathrm{int}(\{K' \in \mathcal{T}_h \mid K \cap K' \ne \emptyset\})$ is a local neighborhood around the mesh cell $K \in \mathcal{T}_h$. We refer the reader to Chapter 22 for a possible construction of the quasi-interpolation operator $\mathcal{I}_{h0}$. (One can also use the canonical interpolation operator defined in §19.4 if r is large enough. In this case, one can use K in lieu of D_K in (57.4).) One should bear in mind that it is not always possible, or easy, to build V_0-conforming finite element subspaces. Think for instance of using Lagrange elements to enforce the value of the normal or tangential component of a vector field at the boundary of a domain that is not a rectangular parallelepiped. We develop in §57.4 a boundary penalty technique that bypasses this difficulty.

57.2.3 Error analysis

We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_{h0} \text{ such that} \\ a^{\mathrm{LS}}(u_h, w_h) = (f, A(w_h))_L, \quad \forall w_h \in V_{h0}. \end{cases} \tag{57.5}$$

Theorem 57.3 (Well-posedness, error bound). (i) *The problem* (57.5) *is well-posed.* (ii) *The following quasi-optimal error bound holds true:*

$$\|u - u_h\|_V \le \frac{\|A\|}{\alpha} \inf_{v_h \in V_{h0}} \|u - v_h\|_V. \tag{57.6}$$

Proof. Well-posedness is a direct consequence of the Lax–Milgram lemma and the V_0-conforming setting. The inequality (57.6) follows from the estimate (26.13) in Céa's lemma since a^{LS} is Hermitian with coercivity and boundedness constants equal to α^2 and $\|A\|^2$, respectively. □

Let $\beta := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(D;\mathbb{C}^{m\times m})}$ and $\phi := \max(\beta, \mu_0 h)$. Assuming $u \in H^{1+r}(D;\mathbb{C}^m)$ and using the approximation properties (57.4) of the quasi-interpolation operator $\mathcal{I}_{h0}$, we infer that

$$\mu_0^{\frac12}\|u - u_h\|_L + \mu_0^{-\frac12}\|A_1(u - u_h)\|_L \le c\,\mu_0^{-\frac12}\phi h^r |u|_{H^{1+r}(D;\mathbb{C}^m)}. \tag{57.7}$$

When $r = k$, the estimate on $\|A_1(u - u_h)\|_L$ has an optimal decay rate w.r.t. $h \in \mathcal{H}$, but this rate is *suboptimal* by one order for $\|u - u_h\|_L$. It is sometimes possible to improve the error estimate in the L-norm by means of the Aubin–Nitsche duality argument, but this is not systematic since, very often, first-order PDEs do not have smoothing properties. This improvement is possible for the one-dimensional transport equation and for Darcy's equation; see Exercises 57.2 and 57.3.

Remark 57.4 (Literature). The LS technique has gained popularity in the numerical analysis community at the beginning of the 1970s following Bramble and Schatz [44, 45], although the technique was already popular in the Russian literature (see Džiškariani [114], Lučka [222]). We refer to Aziz et al. [20] for a theoretical introduction in the context of elliptic problems and to Jiang [198] for a review of applications and implementation aspects. □

Remark 57.5 (Generalizations in $H^{-1}(D)$). One difficulty with the LS technique is that it cannot be extended to H^1-conforming approximations of second-order differential operators. Indeed, if the operator A contains a term such as $-\Delta$, its range is no longer contained in $L^2(D)$ but in $H^{-1}(D)$. As a result, expressions such as $(A(v), A(w))_L$ are no longer meaningful in H^1-conforming spaces. One possible work-around is to use $H^{-1}(D)$ as the pivot space. This strategy is interesting only if a very fast solver (or preconditioner) for the Laplace operator is available. Such solvers (or preconditioners) usually involve a hierarchical decomposition of the approximation space; see Aziz et al. [20], Bramble and Pasciak [43], Bramble et al. [47], Bochev [33], Bramble and Sun [46], Bochev [34], Bramble et al. [48]. □

57.3 Galerkin/least-squares (GaLS)

We consider in this section the GaLS approximation of the problem (57.1).

57.3.1 Local mesh-dependent weights

We define for all $K \in \mathcal{T}_h$ the following local quantities:

$$\beta_K := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(K;\mathbb{C}^{m\times m})}, \tag{57.8}$$

and we introduce the following local weighting parameters:

$$\tau_K := \big(\max(\beta_K h_K^{-1}, \mu_0)\big)^{-1} = \min(\beta_K^{-1} h_K, \mu_0^{-1}), \tag{57.9}$$

where μ_0 comes from Assumption (56.1c). The second equality in (57.9) is meaningful only if β_K is nonzero. We have $\tau_K := \mu_0^{-1}$ if $\beta_K = 0$. For instance, for the advection-reaction equation μ_0 is the reciprocal of a time, β_K is a local velocity, and τ_K is a local time scale. With a slight abuse of notation, we define the piecewise constant function $\tau : D \to \mathbb{R}$ s.t. $\tau_{|K} := \tau_K$ for all $K \in \mathcal{T}_h$. We denote the Euclidean (or Hermitian) norm of $\mathbb{C}^{m\times m}$-valued fields by $\|\cdot\|_{\ell^2}$. Recalling that $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k \mathcal{A}^k$, we assume for simplicity that

$$\max(\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}) \le c_{\mathcal{K},\mathcal{X}}\mu_0, \tag{57.10}$$

and we hide the factor $c_{\mathcal{K},\mathcal{X}}$ in the generic constants used in the error analysis.

57.3.2 Discrete problem and error analysis

We consider the finite element setting of §57.2.2, i.e., we use H^1-conforming finite elements and we strongly enforce the boundary conditions in the discrete setting. We define the following sesquilinear forms on $V_{h0}\times V_{h0}$:

$$a_h^{\mathrm{GL}}(v_h, w_h) := (A(v_h), w_h)_L + r_h(v_h, w_h), \tag{57.11a}$$
$$r_h(v_h, w_h) := (A(v_h), \tau A(w_h))_L. \tag{57.11b}$$

The role of r_h is to stabilize the formulation. Our discrete problem is

$$\begin{cases} \text{Find } u_h \in V_{h0} \text{ such that} \\ a_h^{\mathrm{GL}}(u_h, w_h) = \ell_h^{\mathrm{GL}}(w_h) := (f, w_h + \tau A(w_h))_L, \quad \forall w_h \in V_{h0}. \end{cases} \tag{57.12}$$

Although the approximation setting is conforming, we are in the situation where $a_h^{\mathrm{GL}} \neq a$ and $\ell_h^{\mathrm{GL}} \neq \ell$, i.e., we cannot use the simple setting of §26.3 for the error analysis. Instead, we shall use the more general setting of §27.2. Recall that the three steps of the analysis consist of (i) establishing stability, (ii) bounding the consistency error, and (iii) proving convergence by using the approximation properties of finite elements. Let us start with stability which here takes the simple form of coercivity. Recall the boundary seminorm $|v|_M^2 := \Re\big(\langle M(v), v\rangle_{V',V}\big)$ on V.

Lemma 57.6 (Coercivity, well-posedness). (i) *The following holds true:*

$$\Re\big(a_h^{\mathrm{GL}}(v_h, v_h)\big) \ge \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_M^2 + \|\tau^{\frac12}A(v_h)\|_L^2 =: \|v_h\|_{V_{h0}}^2, \tag{57.13}$$

for all $v_h \in V_{h0}$, i.e., a_h^{GL} is V_{h0}-coercive with constant $\alpha_h := 1$ once V_{h0} is equipped with the norm $\|\cdot\|_{V_{h0}}$. (ii) *The discrete problem* (57.12) *is well-posed.*

Proof. We only need to establish (57.13) since well-posedness then follows from the Lax–Milgram lemma. Using Lemma 56.8 and $V_{h0} \subset V_0 = \ker(M-N)$, we infer that for all $v_h \in V_{h0}$,

$$\begin{aligned}\Re(a_h^{\mathrm{GL}}(v_h,v_h)) &= \Re((A(v_h),v_h)_L) + \|\tau^{\frac12}A(v_h)\|_L^2\\ &\ge \mu_0\|v_h\|_L^2 + \frac12|v_h|_M^2 + \|\tau^{\frac12}A(v_h)\|_L^2 = \|v_h\|_{V_{h0}}^2.\end{aligned}$$

□

The next step consists of bounding the consistency error. Recalling Definition 27.3, the consistency error is defined by

$$\langle \delta_h(v_h), w_h\rangle_{V_{h0}',V_{h0}} := \ell_h^{\mathrm{GL}}(w_h) - a_h^{\mathrm{GL}}(v_h,w_h), \qquad \forall v_h,w_h\in V_{h0}. \tag{57.14}$$

Recalling (27.2) we set $V_{\mathrm{s}} := V_0$ (i.e., no additional smoothness is required on the solution to (57.1)). Owing to the conformity assumption, we have

$$V_\sharp := V_0 + V_{h0} = V_0. \tag{57.15}$$

Contrary to what happens with the nonconforming approximation of elliptic PDEs, we need here to use the setting of Lemma 27.8 which relies on two norms. Specifically, we equip the space $V_\sharp$ with the following two norms:

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac12|v|_M^2 + \|\tau^{\frac12}A(v)\|_L^2, \tag{57.16a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \|\tau^{-\frac12}v\|_L^2. \tag{57.16b}$$

Notice that (27.7) holds true with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_{h0}}$ for all $v_h \in V_{h0}$, and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ for all $v \in V_\sharp = V_0$).

Lemma 57.7 (Consistency/boundedness). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_0$, such that for all $v_h, w_h \in V_{h0}$ and all $h \in \mathcal{H}$:*

$$|\langle \delta_h(v_h), w_h\rangle_{V_{h0}',V_{h0}}| \le \omega_\sharp\, \|u - v_h\|_{V_\sharp}\|w_h\|_{V_{h0}}. \tag{57.17}$$

Proof. Since $A(u) = f$ in L, we have

$$\begin{aligned}\langle \delta_h(v_h), w_h\rangle_{V_{h0}',V_{h0}} &= (f, w_h + \tau A(w_h))_L - (A(v_h),w_h)_L - r_h(v_h,w_h)\\ &= (A(u), w_h+\tau A(w_h))_L - (A(v_h), w_h+\tau A(w_h))_L\\ &= (A(\eta),w_h)_L + (A(\eta),\tau A(w_h))_L,\end{aligned} \tag{57.18}$$

with $\eta := u - v_h$. We bound the two terms on the right-hand side of (57.18). Using integration by parts (see (56.26)), we infer that

$$\begin{aligned}(A(\eta),w_h)_L &= ((\mathcal{K}-\mathcal{X})\eta, w_h)_L - (\eta, A_1(w_h))_L + \langle N(\eta), w_h\rangle_{V',V}\\ &= ((\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})\eta, w_h)_L - (\eta, A(w_h))_L + \langle N(\eta), w_h\rangle_{V',V},\end{aligned}$$

since $A_1(w_h) = A(w_h) - \mathcal{K}w_h$. Let $\mathfrak{T}_1,\mathfrak{T}_2,\mathfrak{T}_3$ be the three terms on the right-hand side. Since $\|\mathcal{K}\|_{\ell^2} = \|\mathcal{K}^{\mathsf{H}}\|_{\ell^2}$, using (57.10) and the Cauchy–Schwarz inequality gives $|\mathfrak{T}_1| \le c\mu_0\|\eta\|_L\|w_h\|_L$. Using again the Cauchy–Schwarz inequality gives $|\mathfrak{T}_2| \le \|\tau^{-\frac12}\eta\|_L\|\tau^{\frac12}A(w_h)\|_L$. Since M is monotone, N is self-adjoint, and $\eta, w_h \in \ker(M-N)$, we infer that $|\mathfrak{T}_3| \le |\eta|_M|w_h|_M$ (see Exercise 56.9). Putting everything together yields $|(A(\eta),w_h)_L| \le c\|\eta\|_{V_\sharp}\|w_h\|_{V_{h0}}$. Finally, using the Cauchy–Schwarz inequality for the second term in (57.18) gives $|(A(\eta),\tau A(w_h))_L| \le \|\tau^{\frac12}A(\eta)\|_L\|\tau^{\frac12}A(w_h)\|_L$. □

Theorem 57.8 (Error estimate). (i) *There is c such that for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\flat} \le c \inf_{v_h \in V_{h0}} \|u - v_h\|_{V_\sharp}. \tag{57.19}$$

(ii) *If $u \in H^{1+r}(D;\mathbb{C}^m)$, $r \in [0,k]$, then*

$$\|u - u_h\|_{V_\flat}^2 \le c \sum_{K\in\mathcal{T}_h} \max(\beta_K, \mu_0 h_K) h_K^{2r+1} |u|^2_{H^{1+r}(D_K;\mathbb{C}^m)}, \tag{57.20}$$

and letting $\phi := \max(\beta, \mu_0 h)$ and $\beta := \max_{K\in\mathcal{T}_h} \beta_K$, we have

$$\|u - u_h\|_{V_\flat} \le c\, \phi_D^{\frac12} h^{r+\frac12} |u|_{H^{1+r}(D;\mathbb{C}^m)}. \tag{57.21}$$

Proof. (i) The error bound (57.19) follows from Lemma 27.8 together with Lemma 57.6 (coercivity) and Lemma 57.7 (consistency/boundedness).

(ii) To prove (57.20), we take $v_h := \mathcal{I}_{h0}(u)$ in (57.19) with $\mathcal{I}_{h0}$ satisfying (57.4). Since $\tau_K^{-\frac12} h_K^{\frac12} = \max(\beta_K^{\frac12}, \mu_0^{\frac12} h_K^{\frac12})$, we infer that for all $K \in \mathcal{T}_h$,

$$\begin{aligned}
\mu_0^{\frac12} \|u - \mathcal{I}_{h0}(u)\|_{L(K)} &\le c\, \mu_0^{\frac12} h_K^{\frac12} h_K^{r+\frac12} |u|_{H^{1+r}(D_K;\mathbb{C}^m)},\\
\|\tau^{-\frac12}(u - \mathcal{I}_{h0}(u))\|_{L(K)} &\le c \max(\beta_K^{\frac12}, \mu_0^{\frac12} h_K^{\frac12}) h_K^{r+\frac12} |u|_{H^{1+r}(D_K;\mathbb{C}^m)}.
\end{aligned}$$

Using (57.10), we infer that $\|\tau^{\frac12} A(v)\|_{L(K)} \le c\tau_K^{\frac12}(\mu_0\|v\|_{L(K)} + \beta_K \|\nabla v\|_{\boldsymbol{L}(K)})$ for all $v \in V$. Hence, taking $v := u - \mathcal{I}_{h0}(u)$ and observing that

$$\tau_K^{\frac12} h_K^{\frac12} (\mu_0 + \beta_K h_K^{-1}) \le \tau_K^{\frac12} h_K^{\frac12} 2\tau_K^{-1} = 2\tau_K^{-\frac12} h_K^{\frac12} = 2\max(\beta_K^{\frac12}, \mu_0^{\frac12} h_K^{\frac12})$$

gives $\|\tau^{\frac12} A(u - \mathcal{I}_{h0}(u))\|_{L(K)} \le c \max(\beta_K^{\frac12}, \mu_0^{\frac12} h_K^{\frac12}) h_K^{r+\frac12} |u|_{H^{1+r}(D_K;\mathbb{C}^m)}$. Regarding the boundary term $|u - \mathcal{I}_{h0}(u)|_M^2$, we observe that $u - \mathcal{I}_{h0}(u) \in V_0$. Using (56.26), we infer that for all $v \in V_0 = \ker(M - N)$,

$$\begin{aligned}
|v|_M^2 = \langle N(v), v\rangle_{V',V} &= (\mathcal{X} v, v)_L + 2\Re((A_1(v), v)_L)\\
&\le c \sum_{K\in\mathcal{T}_h} (\mu_0 \|v\|_{L(K)} + \beta_K \|\nabla v\|_{\boldsymbol{L}(K)}) \|v\|_{L(K)},
\end{aligned}$$

so that $|u - \mathcal{I}_{h0}(u)|_M^2 \le c \sum_{K\in\mathcal{T}_h} \max(\beta_K, \mu_0 h_K) h_K^{2(r+\frac12)} |u|^2_{H^{1+r}(D_K;\mathbb{C}^m)}$. Combining these estimates gives (57.20). Finally, since $\max(\beta_K, \mu_0 h_K) \le \phi$ for all $K \in \mathcal{T}_h$, (57.21) follows from (57.20) because $\sum_{K\in\mathcal{T}_h} |u|^2_{H^{1+r}(D_K;\mathbb{C}^m)} \le c|u|^2_{H^{1+r}(D;\mathbb{C}^m)}$ owing to the regularity of the mesh sequence. □

Assuming $u \in H^{k+1}(D;\mathbb{C}^m)$, the above result implies that

$$\mu_0^{\frac12} \|u - u_h\|_L + \|\tau^{\frac12} A_1(u - u_h)\|_L \le c\, \phi^{\frac12} h^{k+\frac12} |u|_{H^{k+1}(D;\mathbb{C}^m)}.$$

The decay rate of the estimate on $\|u - u_h\|_L$ is improved by half a power in h when compared to that obtained with the LS technique (see (57.7)). Notice also that $A_1(u - u_h)$ is now estimated using a mesh-dependent norm.

Remark 57.9 (Literature). The GaLS technique has been introduced in Hughes et al. [190]. A nonsymmetric variant known under the names of *streamline upwind Petrov–Galerkin* (SUPG) or *streamline diffusion method* has been introduced in Brooks and Hughes [55] and analyzed in Johnson et al. [201]; see Exercise 57.4. We refer the reader to Roos et al. [243, p. 302] for a review of residual-based stabilization techniques. □

57.3.3 Scaling

In applications, the dependent variable u is a vector in $\mathbb{C}^m$ with components each having its own physical dimension. Hence, computing $\|u\|^2_{\ell^2(\mathbb{C}^m)} = \sum_{k\in\{1:m\}} |u_k|^2$ does not make a lot of sense in practice unless the vector u has been made nondimensional from the start. We now address this question.

When the model problem (57.1) is written in dimensional form, we assume that there exists an $m\times m$ symmetric invertible real-valued matrix $\mathcal{S}$ so that the positivity assumption (56.1c) is replaced by

$$((\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})(w),w)_L \geq 2\mu_0\|\mathcal{S}w\|^2_L. \tag{57.22}$$

One can think of $\mathcal{S}w$ as a vector in $\mathbb{C}^m$ whose components all have the same physical dimension. Then the problem (57.1) consists of seeking $v := \mathcal{S}u$ s.t.

$$\mathcal{S}^{-1}\mathcal{K}\mathcal{S}^{-1}v + \sum_{k\in\{1:m\}} \mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1}\partial_{x_k}v = \mathcal{S}^{-1}f. \tag{57.23}$$

Since $\mathcal{S}$ is symmetric, the positivity assumption (57.22) takes the form

$$(\mathcal{S}^{-1}(\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})\mathcal{S}^{-1}(w),w)_L \geq 2\mu_0\|w\|^2_L.$$

Notice also that the matrices $\{\mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1}\}_{k\in\{1:d\}}$ are Hermitian. That is, we recover the theoretical setting discussed in Chapter 56 and in the previous sections of this chapter by replacing $\mathcal{K}$ by $\mathcal{S}^{-1}\mathcal{K}\mathcal{S}^{-1}$ and $\mathcal{A}^k$ by $\mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1}$. We can now write the GaLS formulation of the rescaled problem (57.23). We define for all $K\in\mathcal{T}_h$ the following local rescaled quantities:

$$\beta_K := \max_{k\in\{1:d\}} \|\mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1}\|_{L^\infty(K;\mathbb{C}^{m\times m})}. \tag{57.24}$$

The local weighting parameters τ_K are still defined as in (57.9), where μ_0 now comes from the rescaled positivity assumption (57.22).

Proposition 57.10 (Rescaled GaLS). *Let $v_h\in V_h$ solve the GaLS formulation associated with the rescaled problem* (57.23)*. Then $u_h := \mathcal{S}^{-1}v_h$ solves the following rescaled GaLS formulation: Find $u_h\in V_h$ s.t. for all $z_h\in V_h$,*

$$(A(u_h),z_h)_L + (\mathcal{S}^{-1}A(u_h),\tau\mathcal{S}^{-1}A(z_h))_L = (f,z_h)_L + (\mathcal{S}^{-1}f,\tau\mathcal{S}^{-1}A(z_h))_L.$$

Proof. By definition of v_h, we have for all $w_h\in V_h$,

$$\begin{aligned}(\mathcal{S}^{-1}A(\mathcal{S}^{-1}v_h),w_h)_L + (\mathcal{S}^{-1}A(\mathcal{S}^{-1}v_h),\tau\mathcal{S}^{-1}A(\mathcal{S}^{-1}w_h))_L&\\ = (\mathcal{S}^{-1}f,w_h)_L + (\mathcal{S}^{-1}f,\tau\mathcal{S}^{-1}A(\mathcal{S}^{-1}w_h))_L.&\end{aligned}$$

Setting $u_h := \mathcal{S}^{-1}v_h$ and $z_h := \mathcal{S}^{-1}w_h$, and recalling that $\mathcal{S}$ is symmetric proves the assertion. □

All the error estimates stated in Theorem 57.8 are valid for the rescaled GaLS formulation provided $u-u_h$ and u are replaced in the error estimates by $\mathcal{S}(u-u_h)$ and $\mathcal{S}u$, respectively, and provided v is replaced by $\mathcal{S}u$ and $A(v)$ is replaced by $\mathcal{S}^{-1}A(u)$ in the norms defined in (57.16).

57.3.4 Examples

Let $P_k^{\mathrm{g}}(\mathcal{T}_h)$ be the H^1-conforming finite element subspace defined in §19.2.1 using finite elements of degree $k \geq 1$.

Example 57.11 (Advection-reaction). Consider the PDE $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$ with the inflow boundary condition $u = 0$ on ∂D^-; see §56.2.1. Assume that all the boundary faces of the mesh are subsets of either ∂D^- or $\partial D \backslash \partial D^-$. Let us define $V_{h0} := \{v_h \in P_k^{\mathrm{g}}(\mathcal{T}_h) \mid v_{h|\partial D^-} = 0\}$. By proceeding as in §22.4, a quasi-interpolation operator $\mathcal{I}_{h0}$ can be built by setting to zero the degrees of freedom associated with the boundary faces in ∂D^-. Let us set $\mu_0 := \operatorname{ess\,inf}_D(\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta})$ and define the local weights $\tau_K := \min(\beta_K^{-1} h_K, \mu_0^{-1})$ with $\beta_K := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$. The GaLS discretization consists of seeking $u_h \in V_{h0}$ s.t.

$$\int_D (\mu u_h + \boldsymbol{\beta}\cdot\nabla u_h) w_h \,\mathrm{d}x + \int_D \tau(\mu u_h + \boldsymbol{\beta}\cdot\nabla u_h)(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x = \ell_h^{\mathrm{GL}}(w_h),$$

for all $w_h \in V_{h0}$, with $\ell_h^{\mathrm{GL}}(w_h) := \int_D f w_h \,\mathrm{d}x + \int_D \tau f(\mu w_h + \boldsymbol{\beta}\cdot\nabla w_h)\,\mathrm{d}x$. Let $\beta := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$ and $\phi := \max(\beta, \mu_0 h)$. Assuming that $u \in H^{1+r}(D)$, $r \in [0,k]$, Theorem 57.8 and the approximation properties of V_{h0} give

$$\mu_0^{\frac{1}{2}} \|u - u_h\|_{L^2(D)} + \|\tau^{\frac{1}{2}}\boldsymbol{\beta}\cdot\nabla(u - u_h)\|_{L^2(D)} \leq c\,\phi^{\frac{1}{2}} h^{r+\frac{1}{2}} |u|_{H^{1+r}(D)}. \qquad \square$$

Example 57.12 (Darcy). Consider the PDEs $\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \mathbf{0}$ and $\mu p + \nabla\cdot\boldsymbol{\sigma} = f$ with the boundary condition $p = 0$; see §56.2.2. Notice that $\mathcal{X} = 0$. We are in the situation described in §57.3.3. Let d_* and μ_* be two user-defined reference scales. (Take for instance $d_* := \lambda_\sharp$ and $\mu_* := \mu_\flat$.) Setting $\mu_0 := \min(\frac{\mu_\flat}{\mu_*}, \frac{d_*}{d_\sharp})$, we rewrite (56.16) as follows:

$$((\mathcal{K} + \mathcal{K}^{\mathsf{H}})(\boldsymbol{\sigma}, p), (\boldsymbol{\sigma}, p))_L \geq 2\mu_0 \left(d_*^{-1} \|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 + \mu_* \|p\|_{L^2(D)}^2 \right).$$

The above inequality suggests to consider the following scaling matrix:

$$\mathcal{S} := \left[\begin{array}{c:c} d_*^{-\frac{1}{2}} \mathbb{I}_d & \mathbb{O}_{d\times 1} \\ \hdashline \mathbb{O}_{d\times 1} & \mu_*^{\frac{1}{2}} \end{array}\right]. \tag{57.25}$$

We then observe that $\mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1} = \ell_*\mathcal{A}^k$, where $\ell_* := (d_*/\mu_*)^{\frac{1}{2}}$ is a length scale (for Darcy's equations, the SI unit of d_* is $\mathrm{m}^2\cdot(\mathrm{Pa}\cdot\mathrm{s})^{-1}$ and the SI unit of μ_* is $(\mathrm{Pa}\cdot\mathrm{s})^{-1}$). Then (57.24) implies that $\beta_K := \ell_*$ and the local weighting parameter is $\tau_K := \min(\ell_*^{-1} h_K, \mu_0^{-1})$.

Let us set $\boldsymbol{S}_h := \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$ and $P_{h0} := \{p_h \in P_k^{\mathrm{g}}(\mathcal{T}_h) \mid p_{h|\partial D} = 0\}$. Referring to Proposition 57.10, the rescaled GaLS formulation of the problem consists of seeking $(\boldsymbol{\sigma}_h, p_h) \in V_{h0} := \boldsymbol{S}_h \times P_{h0}$ such that for all $w_h := (\boldsymbol{\tau}_h, q_h) \in V_{h0}$,

$$\begin{aligned}\int_D \Big((\mathbb{d}^{-1}\boldsymbol{\sigma}_h + \nabla p_h)\cdot\boldsymbol{\tau}_h + (\mu p_h + \nabla\cdot\boldsymbol{\sigma}_h) q_h \Big)\,\mathrm{d}x \\ + \int_D d_*\tau(\mathbb{d}^{-1}\boldsymbol{\sigma}_h + \nabla p_h)\cdot(\mathbb{d}^{-1}\boldsymbol{\tau}_h + \nabla q_h)\,\mathrm{d}x \\ + \int_D \mu_*^{-1}\tau(\mu p_h + \nabla\cdot\boldsymbol{\sigma}_h)(\mu q_h + \nabla\cdot\boldsymbol{\tau}_h)\,\mathrm{d}x = \ell_h^{\mathrm{GL}}(w_h),\end{aligned}$$

with $\ell_h^{\mathrm{GL}}(w_h) := \int_D f q_h \,\mathrm{d}x + \int_D \mu_*^{-1}\tau f(\mu q_h + \nabla\cdot\boldsymbol{\tau}_h)\,\mathrm{d}x$. Assume $\boldsymbol{\sigma} \in \boldsymbol{H}^{1+r}(D)$ and $p \in H^{1+r}(D)$, $r \in [0,k]$, and let $\phi := \max(\ell_*, \mu_0 h)$. Then Theorem 57.8 and the approximation properties of V_{h0} give

$$\begin{aligned}&\mu_0^{\frac12}\mu_*^{\frac12}\|p-p_h\|_{L^2(D)} + \mu_0^{\frac12} d_*^{-\frac12}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} + d_*^{\frac12}\|\tau^{\frac12}\nabla(p-p_h)\|_{\boldsymbol{L}^2(D)}\\ &+ \mu_*^{-\frac12}\|\tau^{\frac12}\nabla\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_{L^2(D)} \le c\,\phi^{\frac12}h^{r+\frac12}\big(d_*^{-\frac12}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{1+r}(D)} + \mu_*^{\frac12}|p|_{H^{1+r}(D)}\big).\end{aligned}$$ □

Example 57.13 (Maxwell). Consider the PDEs $\sigma\boldsymbol{E} - \nabla\times\boldsymbol{H} = \boldsymbol{j}_{\mathrm{s}}$ and $\mathrm{i}\omega\mu\boldsymbol{H} + \nabla\times\boldsymbol{E} = \boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$; see §56.2.3. We are in the situation described in §57.3.3. Let σ_* and $\widetilde{\mu}_*$ be two user-defined reference scales. (Take for instance $\sigma_* := \sigma_\flat$ and $\widetilde{\mu}_* := \omega\mu_\flat$.) Setting $\mu_0 := \frac{1}{\sqrt{2}}\min(\frac{\sigma_\flat}{\sigma_*}, \frac{\omega\mu_\flat}{\mu_*})$, we rewrite (56.19) as follows:

$$\Re\big((A(\boldsymbol{E},\boldsymbol{H}),(\boldsymbol{E},\boldsymbol{H}))_L\big) \ge \mu_0\big(\sigma_*\|\boldsymbol{E}\|^2_{\boldsymbol{L}^2(D)} + \widetilde{\mu}_*\|\boldsymbol{H}\|^2_{\boldsymbol{L}^2(D)}\big).$$

The above inequality suggests to consider the following scaling matrix:

$$\mathcal{S} := \left[\begin{array}{c:c} \sigma_*^{\frac12}\mathbb{I}_3 & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & \widetilde{\mu}_*^{\frac12} \end{array}\right]. \tag{57.26}$$

We then observe that $\mathcal{S}^{-1}\mathcal{A}^k\mathcal{S}^{-1} = \ell_*\mathcal{A}^k$, where $\ell_* := (\sigma_*\widetilde{\mu}_*)^{-\frac12}$ is a length scale (for Maxwell's equations, the SI unit of σ_* is $\mathrm{A}^2\cdot\mathrm{s}\cdot(\mathrm{m}\cdot\mathrm{J})^{-1}$ and the SI unit of $\widetilde{\mu}_*$ is $\mathrm{J}\cdot(\mathrm{s}\cdot\mathrm{A}^2\cdot\mathrm{m})^{-1}$). Then (57.24) implies that $\beta_K := \ell_*$ and the local weighting parameter is $\tau_K := \min(\ell_*^{-1}h_K, \mu_0^{-1})$.

Let us set $\boldsymbol{W}_h := \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$ and $\boldsymbol{W}_{h0} := \{\boldsymbol{b}_h \in \boldsymbol{W}_h \mid \boldsymbol{b}_{h|\partial D}\times\boldsymbol{n} = \boldsymbol{0}\}$. Referring to Proposition 57.10, the rescaled GaLS formulation of the problem consists of seeking $(\boldsymbol{E}_h, \boldsymbol{H}_h) \in V_{h0} := \boldsymbol{W}_h\times\boldsymbol{W}_{h0}$ such that

$$\begin{aligned}\int_D \Big((\sigma\boldsymbol{E}_h - \nabla\times\boldsymbol{H}_h)\cdot\overline{\boldsymbol{e}}_h + (\mathrm{i}\omega\mu\boldsymbol{H}_h + \nabla\times\boldsymbol{E}_h)\cdot\overline{\boldsymbol{b}}_h\Big)\,\mathrm{d}x\\ + \int_D \sigma_*^{-1}\tau(\sigma\boldsymbol{E}_h - \nabla\times\boldsymbol{H}_h)\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x\\ + \int_D \tilde{\mu}_*^{-1}\tau(\mathrm{i}\omega\mu\boldsymbol{H}_h + \nabla\times\boldsymbol{E}_h)\cdot(-\mathrm{i}\omega\mu\overline{\boldsymbol{b}}_h + \nabla\times\overline{\boldsymbol{e}}_h)\,\mathrm{d}x = \ell_h^{\mathrm{GL}}(w_h),\end{aligned}$$

for all $w_h := (\boldsymbol{e}_h, \boldsymbol{b}_h) \in V_{h0}$, with $\ell_h^{\mathrm{GL}}(w_h) := \int_D \boldsymbol{j}_{\mathrm{s}}\cdot\overline{\boldsymbol{e}}_h\,\mathrm{d}x + \int_D \sigma_*^{-1}\tau\boldsymbol{j}_{\mathrm{s}}\cdot(\sigma\overline{\boldsymbol{e}}_h - \nabla\times\overline{\boldsymbol{b}}_h)\,\mathrm{d}x$. Assuming that $(\boldsymbol{E},\boldsymbol{H}) \in \boldsymbol{H}^{1+r}(D)\times\boldsymbol{H}^{1+r}(D)$ for some $r \in [0,k]$, Theorem 57.3 combined with the approximation properties of V_{h0} yields

$$\begin{aligned}&\mu_0^{\frac12}\sigma_*^{\frac12}\|\boldsymbol{E}-\boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)} + \mu_0^{\frac12}\tilde{\mu}_*^{\frac12}\|\boldsymbol{H}-\boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)} + \tilde{\mu}_*^{-\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{E}-\boldsymbol{E}_h)\|_{\boldsymbol{L}^2(D)}\\ &\qquad + \sigma_*^{-\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{H}-\boldsymbol{H}_h)\|_{\boldsymbol{L}^2(D)} \le c\,\phi^{\frac12}h^{r+\frac12}\big(\sigma_*^{\frac12}|\boldsymbol{E}|_{\boldsymbol{H}^{1+r}(D)} + \widetilde{\mu}_*^{\frac12}|\boldsymbol{H}|_{\boldsymbol{H}^{1+r}(D)}\big),\end{aligned}$$

with $\phi := \max(\ell_*, \mu_0 h)$. If the boundary of D is not smooth and/or if the coefficients σ, μ are discontinuous, it is in general preferable to use $\boldsymbol{H}(\mathrm{curl})$-conforming finite element subspaces based on edge elements (see Chapter 15) instead of using H^1-conforming finite element subspaces; see §45.4. □

57.4 Boundary penalty for Friedrichs' systems

The goal of this section is twofold. First, we introduce a technique to enforce boundary conditions weakly in Friedrichs' systems. Then we show how to combine this technique with the GaLS method. The boundary penalty technique introduced here will be used again in the following chapters.

57.4.1 Model problem

Recalling that $V_0 := \ker(M-N)$, we consider the sesquilinear form

$$\widetilde{a}(v,w) := (A(v),w)_L + \frac{1}{2}\langle (M-N)(v), w\rangle_{V',V}, \qquad \forall v,w \in V. \tag{57.27}$$

The purpose of the last term on the right-hand side is to enforce the boundary condition $u \in \ker(M-N)$ weakly. The test functions are now restricted to be in the graph space V, since the linear form $(M-N)(v)$ is not bounded on L. The model problem that we consider is the following:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \widetilde{a}(u,w) = (f,w)_L, \quad \forall w \in V. \end{cases} \tag{57.28}$$

If u solves (57.28), taking w in $C_0^\infty(D;\mathbb{C}^m)$ implies that $A(u) = f$ in $L^2(D;\mathbb{C}^m)$. Then we have $\langle (M-N)(u), w\rangle_{V',V} = 0$ for all $w \in V$, which implies that $u \in \ker(M-N)$.

Lemma 57.14 (L-coercivity). *The sesquilinear form $\widetilde{a}$ defined in* (57.27) *has the following coercivity property:*

$$\Re\big(\widetilde{a}(v,v)\big) \geq \mu_0 \|v\|_L^2 + \frac{1}{2}|v|_M^2, \qquad \forall v \in V. \tag{57.29}$$

Proof. Owing to Lemma 56.8, we infer that

$$\begin{aligned} \Re\big(\widetilde{a}(v,v)\big) &= \Re\big((A(v),v)_L\big) + \frac{1}{2}\Re\big(\langle (M-N)(v), v\rangle_{V',V}\big) \\ &\geq \mu_0\|v\|_L^2 + \frac{1}{2}\Re\big(\langle N(v), v\rangle_{V',V}\big) + \frac{1}{2}\Re\big(\langle (M-N)(v), v\rangle_{V',V}\big), \end{aligned}$$

for all $v \in V$, so that (57.29) follows readily. □

Proposition 57.15 (Well-posedness). (i) *The problem* (57.28) *is well-posed.* (ii) *Its unique solution is the unique solution to* (57.1).

Proof. Assume that u solves (57.1). Then $u \in V_0 = \ker(M-N)$ and we have $(A(u),w)_L = (f,w)_L$ for all $w \in L$. Hence, $\widetilde{a}(u,w) = (A(u),w)_L = (f,w)_L$ for all $w \in V \subset L$. This shows that u solves (57.28). The uniqueness of the solution to (57.28) results from Lemma 57.14 and $\mu_0 > 0$. □

Remark 57.16 (Inf-sup condition). One should not infer from the well-posedness of (57.28) that the sesquilinear form $\widetilde{a}$ satisfies an inf-sup condition on $V \times V$. Indeed, well-posedness holds true for all $f \in L$, but it may not be the case for all $f \in V'$. □

57.4.2 Boundary penalty method

We now construct a V-conforming approximation of the model problem (57.28) by using H^1-conforming finite elements. We denote by $(V_h)_{h\in\mathcal{H}}$ a sequence of H^1-conforming finite element subspaces constructed as in §57.2.2 using a reference finite element of degree $k \geq 1$. We assume

that we have at hand a quasi-interpolation operator $\mathcal{I}_h : V \to V_h$ with optimal local approximation properties, i.e., there is c s.t. for all $r \in [0,k]$, all $v \in H^{1+r}(D;\mathbb{C}^m)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,

$$\|v - \mathcal{I}_h(v)\|_{L(K)} + h_K \|\nabla(v - \mathcal{I}_h(v))\|_{\boldsymbol{L}(K)} \le c\, h_K^{1+r} |v|_{H^{1+r}(D_K,\mathbb{C}^m)}. \tag{57.30}$$

Our starting point is the sesquilinear form $\widetilde{a}$ defined in (57.27). We would like to localize the term $\langle (M-N)(v), w\rangle_{V',V}$ over the boundary faces $F \in \mathcal{F}_h^\partial$. To this end, we assume that there are boundary fields $\mathcal{M},\mathcal{N} \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ s.t.

$$\langle M(v), w\rangle_{V',V} = (\mathcal{M}v, w)_{L(\partial D)}, \quad \langle N(v), w\rangle_{V',V} = (\mathcal{N}v, w)_{L(\partial D)}, \tag{57.31}$$

for all $v, w \in H^s(D;\mathbb{C}^m)$, $s > \frac{1}{2}$, and $L(\partial D) := L^2(\partial D;\mathbb{C}^m)$. Notice that the field $\mathcal{M}$ is such that $\Re\big((\mathcal{M}v, v)_{L(\partial D)}\big) \ge 0$ since the operator M is monotone. But the examples from §56.2 show that $\Re\big((\mathcal{M}v, v)_{L(\partial D)}\big)$ may vanish identically (this happens for second-order PDEs in mixed form). To gain some control on the boundary values, we thus need to introduce an additional boundary penalty field $\mathcal{S}^\partial \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ taking values in the set of $m\times m$ complex-valued matrices that are Hermitian and positive semidefinite. We define the following seminorm on $H^s(D;\mathbb{C}^m)$, $s > \frac{1}{2}$:

$$|v|_{\mathcal{M}^{\mathrm{BP}}} := (\mathcal{M}^{\mathrm{BP}} v, v)^{\frac{1}{2}}_{L(\partial D)}, \qquad \mathcal{M}^{\mathrm{BP}} := \mathcal{M} + \mathcal{S}^\partial. \tag{57.32}$$

Letting $L(F) := L^2(F;\mathbb{C}^m)$, we define the seminorm $|v|_{\mathcal{M}_F^{\mathrm{BP}}} = (\mathcal{M}_F^{\mathrm{BP}} v, v)^{\frac{1}{2}}_{L(F)}$, where we use the subscript $_F$ for the restriction of a boundary field to $F \in \mathcal{F}_h^\partial$. We assume that the field $\mathcal{S}_F^\partial$ is defined in such a way that there is c s.t.

$$\ker(\mathcal{M}_F - \mathcal{N}_F) \subset \ker(\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F), \tag{57.33a}$$

$$|v|_{\mathcal{M}_F^{\mathrm{BP}}} \le c\, \beta_{K_l}^{\frac{1}{2}} \|v\|_{L(F)}, \tag{57.33b}$$

$$|((\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)v, w)_{L(F)}| \le c\, \beta_{K_l}^{\frac{1}{2}} |v|_{\mathcal{M}_F^{\mathrm{BP}}} \|w\|_{L(F)}, \tag{57.33c}$$

$$|((\mathcal{M}_F^{\mathrm{BP}} + \mathcal{N}_F)v, w)_{L(F)}| \le c\, \beta_{K_l}^{\frac{1}{2}} \|v\|_{L(F)} |w|_{\mathcal{M}_F^{\mathrm{BP}}}, \tag{57.33d}$$

for all $v, w \in L(F)$, all $F \in \mathcal{F}_h^\partial$, and all $h \in \mathcal{H}$, where β_K is defined in (57.8) for all $K \in \mathcal{T}_h$, and K_l is the unique mesh cell of which F is a face, i.e., $F := \partial K_l \cap \partial D$. Notice that (57.33c) and (57.33d) turn out to be equivalent (see Exercise 57.5). We retain both (57.33c) and (57.33d) as assumptions since each will appear in the analysis and this avoids distracting technicalities.

Example 57.17 (Advection-reaction). Since $\mathcal{M}_F = |\boldsymbol{\beta}\cdot\boldsymbol{n}_F|$ for all $F \in \mathcal{F}_h^\partial$, we can take $\mathcal{S}_F^\partial := 0$. The properties (57.33a) and (57.33b) are obvious, and (57.33c) results from the Cauchy–Schwarz inequality since $\frac{1}{2}\int_F (|\boldsymbol{\beta}\cdot\boldsymbol{n}_F| - \boldsymbol{\beta}\cdot\boldsymbol{n}_F) vw \,\mathrm{d}s \le \||\boldsymbol{\beta}\cdot\boldsymbol{n}_F|^{\frac{1}{2}} v\|_{L^2(F)} \beta_{K_l}^{\frac{1}{2}} \|w\|_{L^2(F)}$. □

Example 57.18 (Darcy). The properties (57.33) are satisfied for the Dirichlet condition $p = 0$ by taking

$$\mathcal{S}_F^\partial := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & 0 \\ \hdashline 0 & \alpha_F \end{array}\right], \qquad \forall F \in \mathcal{F}_h^\partial. \tag{57.34}$$

Recalling the scaling arguments from §57.3.3 and Example 57.12, the inequality (57.33b) requires that $(\alpha_F p, p)^{\frac{1}{2}}_{L^2(F)} \le c \beta_{K_l}^{\frac{1}{2}} \mu_*^{\frac{1}{2}} \|p\|_{L^2(F)}$. We then set $\alpha_F := \alpha_* \beta_{K_l} \mu_*$, where $\beta_{K_l} := \ell_* := (d_*/\mu_*)^{\frac{1}{2}}$ is a length scale and $\alpha_* > 0$ is a user-defined $\mathcal{O}(1)$ nondimensional parameter. □

Example 57.19 (Maxwell). The properties (57.33) are satisfied for the boundary condition $\boldsymbol{H}\times\boldsymbol{n}=\boldsymbol{0}$ by taking

$$\mathcal{S}_F^\partial := \left[\begin{array}{c:c} \mathbb{O}_{3\times3} & \mathbb{O}_{3\times3} \\ \hdashline \mathbb{O}_{3\times3} & \alpha_F\mathbb{T}^\intercal\mathbb{T} \end{array}\right], \qquad \forall F\in\mathcal{F}_h^\partial, \tag{57.35}$$

where $\mathbb{T}\in\mathbb{R}^{3\times3}$ is s.t. $\mathbb{T}\boldsymbol{\xi} := \boldsymbol{\xi}\times\boldsymbol{n}$ for all $\boldsymbol{\xi}\in\mathbb{R}^3$ (see §56.2.3). Recalling the scaling arguments from §57.3.3 and Example 57.13, the inequality (57.33b) requires that $\|\alpha_F^{\frac12}\boldsymbol{H}\times\boldsymbol{n}\|_{L^2(F)}\le c\beta_{K_l}^{\frac12}\widetilde{\mu}_*^{\frac12}\|\boldsymbol{H}\|_{\boldsymbol{L}^2(F)}$. We then set $\alpha:=\alpha_*\beta_{K_l}\widetilde{\mu}_*$, where $\beta_{K_l}:=\ell_*:=(d_*\widetilde{\mu}_*)^{-\frac12}$ is a length scale and $\alpha_*>0$ is a user-defined $\mathcal{O}(1)$ nondimensional parameter. □

57.4.3 GaLS stabilization with boundary penalty

Let us now formulate the GaLS approximation with boundary penalty. Recalling that $r_h(v_h,w_h):=(A(v_h),\tau A(w_h))_L$, we define the following discrete sesquilinear forms on $V_h\times V_h$:

$$a_h^{\text{BP}}(v_h,w_h) := (A(v_h),w_h)_L + \frac12((\mathcal{M}^{\text{BP}}-\mathcal{N})v_h,w_h)_{L(\partial D)}, \tag{57.36a}$$

$$a_h^{\text{GL/BP}}(v_h,w_h) := a_h^{\text{BP}}(v_h,w_h) + r_h(v_h,w_h). \tag{57.36b}$$

We consider the following discrete problem:

$$\left\{\begin{array}{l} \text{Find } u_h\in V_h \text{ such that} \\ a_h^{\text{GL/BP}}(u_h,w_h) = \ell_h^{\text{GL}}(w_h) := (f, w_h+\tau A(w_h))_L, \quad \forall w_h\in V_h. \end{array}\right. \tag{57.37}$$

Note that the boundary penalty technique does not affect the definition of the right-hand side, unless the boundary condition is non-homogeneous. We perform the error analysis as in §57.3.2 using Lemma 27.8. Let us start with stability which again takes the simple form of coercivity.

Lemma 57.20 (Coercivity, well-posedness). (i) *The following holds true:*

$$\Re\big(a_h^{\text{GL/BP}}(v_h,v_h)\big) \ge \mu_0\|v_h\|_L^2 + \frac12|v_h|_{\mathcal{M}^{\text{BP}}}^2 + \|\tau^{\frac12}A(v_h)\|_L^2 =: \|v_h\|_{V_h}^2, \tag{57.38}$$

for all $v_h\in V_h$. (ii) *The discrete problem* (57.37) *is well-posed.*

Proof. Similar to that of Lemma 57.6. □

Since the boundary penalty technique invokes traces on ∂D, some additional smoothness on the solution has to be assumed, i.e., we assume that

$$u\in V_{\text{S}} := H^s(D;\mathbb{C}^m)\cap V, \qquad s>\frac12. \tag{57.39}$$

We set $V_\sharp := V_{\text{S}}+V_h = V_{\text{S}}$ (owing to conformity), and we equip the space $V_\sharp$ with the following two norms:

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac12|v|_{\mathcal{M}^{\text{BP}}}^2 + \|\tau^{\frac12}A(v)\|_L^2, \tag{57.40a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \|\tau^{-\frac12}v\|_L^2 + \|\rho^{\frac12}v\|_{L(\partial D)}^2, \tag{57.40b}$$

with the scaling factor $\rho\in L^\infty(\partial D)$ defined by setting $\rho_{|F}:=\beta_{K_l}$ for all $F\in\mathcal{F}_h^\partial$. Notice that $\|v_h\|_{V_h}=\|v_h\|_{V_\flat}$ for all $v_h\in V_h$ and $\|v\|_{V_\flat}\le\|v\|_{V_\sharp}$ for all $v\in V_\sharp$, i.e., (27.7) holds true with $c_\flat:=1$.

Lemma 57.21 (Consistency/boundedness). *Define the consistency error as*

$$\langle \delta_h(v_h), w_h \rangle_{V_h',V_h} := \ell_h^{\mathrm{GL}}(w_h) - a_h^{\mathrm{GL/BP}}(v_h, w_h), \qquad \forall v_h, w_h \in V_h.$$

There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{S}}$, such that for all $v_h, w_h \in V_h$ and all $h \in \mathcal{H}$,

$$|\langle \delta_h(v_h), w_h \rangle_{V_h',V_h}| \le \omega_\sharp \, \|u - v_h\|_{V_\sharp} \|w_h\|_{V_h}. \tag{57.41}$$

Proof. Since $A(u) = f$ in L, $(\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)u = 0$ in $L(\partial D)$ owing to (57.33a), and since we assumed $u \in H^s(D;\mathbb{C}^m)$, $s > \frac{1}{2}$, we have

$$\begin{aligned}\langle \delta_h(v_h), w_h \rangle_{V_h',V_h} &= (A(u), w_h + \tau A(w_h))_L - (A(v_h), w_h)_L \\ &\quad - \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})v_h, w_h)_{L(\partial D)} - (A(v_h), \tau A(w_h))_L \\ &= (A(\eta), w_h + \tau A(w_h))_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})\eta, w_h)_{L(\partial D)},\end{aligned}$$

with $\eta := u - v_h$. Integrating by parts, we obtain

$$\begin{aligned}&(A(\eta), w_h)_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})\eta, w_h)_{L(\partial D)} \\ &\qquad = ((\mathcal{K} + \mathcal{K}^{\mathsf{H}} - \mathcal{X})\eta, w_h)_L - (\eta, A(w_h))_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta, w_h)_{L(\partial D)}.\end{aligned}$$

The third term on the right-hand side is bounded by using (57.33d). The rest of the proof is similar to that of Lemma 57.7. □

Theorem 57.22 (Error estimate). (i) *There is c such that for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\flat} \le c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{57.42}$$

(ii) *If $u \in H^{1+r}(D;\mathbb{C}^m)$, $r \in [0,k]$, then*

$$\|u - u_h\|_{V_\flat}^2 \le c \sum_{K \in \mathcal{T}_h} \max(\beta_K, \mu_0 h_K) h_K^{2r+1} |u|_{H^{1+r}(D_K;\mathbb{C}^m)}^2. \tag{57.43}$$

This implies in particular that $\|u - u_h\|_{V_\flat} \le c\phi_D^{\frac{1}{2}} h^{r+\frac{1}{2}} |u|_{H^{1+r}(D;\mathbb{C}^m)}$.

Proof. Similar to that of Theorem 57.8. □

Exercises

Exercise 57.1 (Least-squares). Write the LS approximation and the resulting error estimate for the advection-reaction, Darcy's, and Maxwell's equations (for simplicity assume that $u \in H^{k+1}(D;\mathbb{C}^m)$ and hide the scaling factors in the generic constant c).

Exercise 57.2 (Transport in 1D). Consider the LS approximation using $\mathbb{P}_k$ Lagrange finite elements, $k \ge 1$, of the one-dimensional transport problem $u' = f$ in $D := (0,1)$ with $u(0) = 0$ and $f \in H^k(D)$. Prove the optimal L^2-error estimate $\|u - u_h\|_{L^2(D)} \le ch^{k+1}|f|_{H^k(D)}$. (*Hint*: use a duality argument.)

Exercise 57.3 (Duality argument for Darcy). Consider the LS approximation of Darcy's equations with homogeneous Dirichlet conditions on p in the mixed-order case $k := k_{\boldsymbol{\sigma}} - 1 = k_p \geq 1$, i.e., $V_{h0} := \boldsymbol{P}^{\mathrm{g}}_{k+1}(\mathcal{T}_h) \times P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$. Assume that $\mu := 0$, $\mathbb{d}^{-1} := \kappa \mathbb{I}_d$ with $\kappa \in W^{1,\infty}(D)$, and that full elliptic regularity holds true for the Laplacian. The goal is to prove the error bound $\|p - p_h\|_{L^2(D)} \leq ch^{k+1}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)} + |p|_{H^{k+1}(D)})$; see Pehlivanov et al. [236]. Let $\mathcal{I}_h$ have optimal approximation properties in $\boldsymbol{P}^{\mathrm{g}}_{k+1}(\mathcal{T}_h)$, and let $\Pi^{\mathrm{e}}_h : H^1_0(D) \to P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$ be the elliptic projection such that for all $q \in H^1_0(D)$, $(\nabla(q - \Pi^{\mathrm{e}}_h(q)), \nabla q_h)_{\boldsymbol{L}^2(D)} = 0$ for all $q_h \in P^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$ (see §32.4). (i) Setting $e_h := (\mathcal{I}_h(\boldsymbol{\sigma}) - \boldsymbol{\sigma}_h, \Pi^{\mathrm{e}}_h(p) - p_h)$, prove that $\|e_h\|_V \leq c(\|\mathcal{I}_h(\boldsymbol{\sigma}) - \boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)} + \|\Pi^{\mathrm{e}}_h(p) - p\|_{L^2(D)})$. (*Hint*: use coercivity and the Galerkin orthogonality property.) (ii) Show that $\|p - p_h\|_{L^2(D)} \leq ch^{k+1}(|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+2}(D)} + |p|_{H^{k+1}(D)})$. (*Hint*: use a Poincaré–Steklov inequality and Exercise 32.1.)

Exercise 57.4 (SUPG). Assume that $h_K \leq \beta_K \mu_0^{-1} \min(1, \frac{1}{2}\frac{\mu_0^2}{\mu_\infty^2})$ for all $K \in \mathcal{T}_h$ with $\mu_\infty := \|\mathcal{K}\|_{\mathbb{L}^\infty(D)}$. Prove that the same error estimate as in the GaLS approximation is obtained by considering the following discrete problem: Find $u_h \in V_{h0}$ such that $a_h^{\mathrm{SUPG}}(u_h, w_h) = (f, w_h + \tau A_1(w_h))_L$ for all $w_h \in V_{h0}$ with the SUPG-stabilized sesquilinear form $a_h^{\mathrm{SUPG}}(v_h, w_h) := (A(v_h), w_h)_L + (A(v_h), \tau A_1(w_h))_L$. (*Hint*: bound $(A(v_h), \tau \mathcal{K} v_h)_L$ and use Lemma 57.6 to establish coercivity.)

Exercise 57.5 (Boundary penalty). (i) Prove that (57.33c) and (57.33d) are equivalent. (*Hint*: consider the Hermitian and skew-Hermitian parts of $\mathcal{M}_F$.) (ii) Verify that the boundary penalty operators defined in Example 57.18 for Darcy's equations and in Example 57.19 for Maxwell's equations satisfy (57.33). (*Hint*: direct verification.)

Chapter 58

Fluctuation-based stabilization (I)

The goal of this chapter and the next one is to approximate the same model problem as in Chapter 57, still with H^1-conforming finite elements and the boundary penalty technique introduced in §57.4, but with a different stabilization technique. One motivation is that the residual-based stabilization is delicate to use when approximating time-dependent PDEs since the time derivative is part of the residual. The techniques devised in this chapter and the next one avoid this difficulty. The starting observation is that H^1-conforming test functions cannot control the gradient of H^1-conforming functions since the gradient generally exhibits jumps across the mesh interfaces. The idea behind fluctuation-based stabilization is to gain full control on the gradient by adding a least-squares penalty on the part of the gradient departing from the H^1-conforming space, and this part can be viewed as a fluctuation. Stabilization techniques based on this idea include the continuous interior penalty (CIP) method, studied in this chapter, and two-scale stabilization techniques such as the local projection stabilization (LPS) and the subgrid viscosity (SGV) methods, which are studied in the next chapter. We present in this chapter a unified analysis based on an abstract set of assumptions. We show how to satisfy these assumptions using CIP, LPS, and SGV in this chapter and the next one. Notice that in terms of stability and approximation, GaLS, CIP, LPS, SGV, and discontinuous Galerkin (presented in Chapter 60) are all equivalent.

58.1 Discrete setting

We assume that for all $h \in \mathcal{H}$, we have at hand an H^1-conforming finite-dimensional space $V_h \subset V$ built by using a shape-regular mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ and a finite element of degree $k \geq 1$. We assume that each mesh covers D exactly. We also assume that there is a quasi-interpolation operator $\mathcal{I}_h : V \to V_h$ with optimal local approximation properties, i.e., there is c s.t.

$$|v - \mathcal{I}_h(v)|_{H^l(K;\mathbb{C}^m)} \leq c\, h_K^{1+r-l} |v|_{H^{1+r}(D_K,\mathbb{C}^m)}, \tag{58.1}$$

for all $r \in [0, k]$, every integer $l \in \{0{:}1{+}\lfloor r \rfloor\}$, all $v \in H^{1+r}(D;\mathbb{C}^m)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$, where $D_K := \mathrm{int}(\{K' \in \mathcal{T}_h \mid K \cap K' \neq \emptyset\})$.

Recall from §57.3.1 the local quantities $\beta_K := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(K;\mathbb{C}^{m\times m})}$ for all $K \in \mathcal{T}_h$, and the local weighting parameters

$$\tau_K := \big(\max(\beta_K h_K^{-1}, \mu_0)\big)^{-1} = \min(\beta_K^{-1} h_K, \mu_0^{-1}), \tag{58.2}$$

where μ_0 is defined in (56.1c). For the advection-reaction equation for instance, β_K is a local velocity scale, μ_0 is the reciprocal of a time, and τ_K is a local time scale. We define the global quantity $\beta := \max_{K\in\mathcal{T}_h}\beta_K$. With a slight abuse of notation, we define the piecewise constant function $\tau : D \to \mathbb{R}$ s.t. $\tau_{|K} := \tau_K$ for all $K \in \mathcal{T}_h$.

Our starting point is the boundary penalty technique that we used for the GaLS stabilization (see (57.36a)). Recall that $L := L^2(D;\mathbb{C}^m)$ and $L(\partial D) := L^2(\partial D;\mathbb{C}^m)$. The sesquilinear form is defined by setting

$$a_h^{\mathrm{BP}}(v_h,w_h) := (A(v_h),w_h)_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}}-\mathcal{N})v_h,w_h)_{L(\partial D)}, \tag{58.3}$$

for all $v_h, w_h \in V_h$. The main idea of the fluctuation-based stabilization is to augment a_h^{BP} with a Hermitian positive semidefinite sesquilinear form s_h whose purpose is loosely speaking to control the difference between $A_1(v_h)$ and a suitable representative of $A_1(v_h)$ in V_h. We make the following requirements on s_h: There exists a linear operator $\mathcal{J}_h : V_h \to V_h$ and two positive constants c_1, c_2 s.t. the following holds true for all $v_h \in V_h$ and all $h \in \mathcal{H}$:

$$|v_h|_{\mathcal{S}} := s_h(v_h,v_h)^{\frac12} \le c_1\|\tau^{-\frac12}v_h\|_L, \tag{58.4a}$$

$$c_2\,\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L^2 \le \|\tau^{\frac12}A_1(v_h)\|_L^2 + \mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2, \tag{58.4b}$$

$$c_2\,\|\tau^{\frac12}A_1(v_h)\|_L^2 \le \Re\big((A_1(v_h),\mathcal{J}_h(v_h))_L\big) + \mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2. \tag{58.4c}$$

We are going to give examples for s_h and $\mathcal{J}_h$ in §58.3 for CIP and in the next chapter for LPS and SGV. With this new tool in hand, we consider the following discrete problem:

$$\left\{\begin{array}{l}\text{Find } u_h \in V_h \text{ such that}\\ a_h^{\mathrm{F1}}(u_h,w_h) = \ell_h^{\mathrm{F1}}(w_h), \quad \forall w_h \in V_h,\end{array}\right. \tag{58.5}$$

with $a_h^{\mathrm{F1}}(v_h,w_h) := a_h^{\mathrm{BP}}(v_h,w_h) + s_h(v_h,w_h)$ and $\ell_h^{\mathrm{F1}}(w_h) := (f,w_h)_L$. Notice that the right-hand side ℓ_h^{F1} does not depend on the stabilization.

Remark 58.1 (Simplified setting). Let $\ell_D := \mathrm{diam}(D)$ and assume that $\mu_0 \ge \beta\ell_D^{-1}$ and that the mesh family $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is quasi-uniform. Then one can consider the constant coefficient $\tau := \beta^{-1}h$ and (58.4a) becomes

$$|v_h|_{\mathcal{S}} \le c_1\big(\tfrac{\beta}{h}\big)^{\frac12}\|v_h\|_L, \qquad \forall v_h \in V_h. \tag{58.6}$$

Moreover, if one can devise a linear operator $\mathcal{A}_h : V_h \to V_h$ s.t. for all $v_h \in V_h$,

$$\|A_1(v_h) - \mathcal{A}_h(v_h)\|_L \le c\,\Big(\big(\tfrac{\beta}{h}\big)^{\frac12}|v_h|_{\mathcal{S}} + \tfrac{\beta}{\ell_D}\|v_h\|_L\Big), \tag{58.7}$$

then (58.4b)-(58.4c) are met with $\mathcal{J}_h(v_h) := \frac{h}{\beta}\mathcal{A}_h(v_h)$; see Exercise 58.1. Let $\mathcal{P}_{V_h} : L \to V_h$ be the L-orthogonal projection onto V_h, i.e., for all $z \in L$, $\mathcal{P}_{V_h}(z) \in V_h$ is uniquely defined s.t. $(z-\mathcal{P}_{V_h}(z),w_h)_L = 0$ for all $w_h \in V_h$. Then under the above assumptions, all the fluctuation-based sesquilinear forms s_h (whether CIP, LPS, or SGV based) lead to the following decay rates: There is c s.t. for all $v \in H^{k+1}(D;\mathbb{C}^m)$, all $w_h \in V_h$, and all $h \in \mathcal{H}$,

$$|s_h(\mathcal{P}_{V_h}(v),w_h)| \le c\,\beta^{\frac12}h^{k+\frac12}|v|_{H^{k+1}(D;\mathbb{C}^m)}|w_h|_{\mathcal{S}}, \tag{58.8a}$$

$$\begin{aligned}|(v-\mathcal{P}_{V_h}(v),A_1(w_h))_L| &\le c\,\beta^{\frac12}h^{k+\frac12}|v|_{H^{k+1}(D;\mathbb{C}^m)}\\ &\quad\times\Big(|w_h|_{\mathcal{S}} + \big(\tfrac{\beta}{\ell_D}\big)^{\frac12}\|w_h\|_L\Big).\end{aligned} \tag{58.8b}$$

□

58.2 Stability analysis

The goal of this section is to prove that the discrete sesquilinear form a_h^{Fl} defined above satisfies an inf-sup condition on $V_h \times V_h$ uniformly w.r.t. $h \in \mathcal{H}$. We assume that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are piecewise Lipschitz on a partition of D and that the meshes are compatible with this partition, implying that the fields $\{\mathcal{A}^k_{|K}\}_{k\in\{1:d\}}$ are Lipschitz for all $K \in \mathcal{T}_h$. We denote by $L_{\mathcal{A}}$ the largest Lipschitz constant of these fields. Recall that $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k \mathcal{A}^k$. To simplify the tracking of the model parameters, we assume that

$$\max(\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, L_{\mathcal{A}}) \le c_{\mathcal{K},\mathcal{X},\mathcal{A}}\mu_0, \tag{58.9}$$

and we hide $c_{\mathcal{K},\mathcal{X},\mathcal{A}}$ in the generic constants c used in the error analysis. Notice that we have $\|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})} \le dL_{\mathcal{A}}$.

Lemma 58.2 (Inf-sup stability, well-posedness). (i) *Under the conditions* (58.4) *on* s_h *and the above assumption on the model parameters, there is* $\alpha > 0$ *such that for all* $h \in \mathcal{H}$,

$$\inf_{v_h\in V_h} \sup_{w_h\in V_h} \frac{|a_h^{\mathrm{Fl}}(v_h, w_h)|}{\|v_h\|_{V_h}\|w_h\|_{V_h}} \ge \alpha > 0, \tag{58.10}$$

with the stability norm

$$\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |v_h|_{\mathcal{S}}^2 + \|\tau^{\frac12}A_1(v_h)\|_L^2. \tag{58.11}$$

(ii) *The discrete problem* (58.5) *is well-posed.*

Proof. We only need to prove (58.10) since the well-posedness of (58.5) then follows directly. Let $v_h \in V_h$. Set $l_h := \|v_h\|_{V_h}$ and $r_h := \sup_{w_h\in V_h} \frac{|a_h^{\mathrm{Fl}}(v_h,w_h)|}{\|w_h\|_{V_h}}$. Our goal is to prove that there is $\alpha > 0$ such that $\alpha l_h \le r_h$ for all $h \in \mathcal{H}$.

(1) The coercivity of a_h^{BP} and the positive semidefiniteness of s_h imply that

$$\mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |v_h|_{\mathcal{S}}^2 \le \Re\big(a_h^{\mathrm{Fl}}(v_h, v_h)\big) \le r_h l_h. \tag{58.12}$$

(2) Let us set $w_h := \mathcal{J}_h(v_h)$ and let us verify that $\|w_h\|_{V_h} \le c\|\tau^{-\frac12}w_h\|_L$. To this purpose, we bound the four terms composing $\|w_h\|_{V_h}$. For the first term, $\mu_0^{\frac12}\|w_h\|_L$, we use the fact that $\mu_0 \le \tau_K^{-1}$ for all $K \in \mathcal{T}_h$. For the second term, we use (57.33b), a discrete trace inequality, and the fact that $\beta_{K_l}h_{K_l}^{-1} \le \tau_{K_l}^{-1}$. For the third term, we use the property $|w_h|_{\mathcal{S}} \le c_1\|\tau^{-\frac12}w_h\|_L$ from the design condition (58.4a) on s_h. For the fourth term, $\|\tau^{\frac12}A_1(w_h)\|_L$, we use the definition of β_K, an inverse inequality, and that $\beta_K h_K^{-1} \le \tau_K^{-1}$. Putting together the above bounds shows that $\|w_h\|_{V_h} \le c\|\tau^{-\frac12}w_h\|_L$. Owing to (58.4b) and (58.12), and recalling that $l_h := \|v_h\|_h$, we infer that

$$\|w_h\|_{V_h} \le c\,\|\tau^{-\frac12}w_h\|_L = c\,\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L \le c\,c_2^{-\frac12}\,l_h. \tag{58.13}$$

(3) Using (58.12), the condition (58.4c) implies that

$$\|\tau^{\frac12}A_1(v_h)\|_L^2 \le c\big(\Re\big((A_1(v_h), w_h)_L\big) + r_h l_h\big). \tag{58.14}$$

Summing (58.12) and (58.14) (which amounts to using the test function $v_h + \mathcal{J}_h(v_h)$) gives $l_h^2 \le c(\Re((A_1(v_h), w_h)_L) + r_h l_h)$. The rest of the proof consists of estimating $\Re((A_1(v_h), w_h)_L)$. This term is rewritten as follows:

$$\Re\big((A_1(v_h), w_h)_L\big) = \Re\big(a_h^{\mathrm{Fl}}(v_h, w_h)\big) - \Upsilon_h,$$

where $\Upsilon_h := \Re\big((\mathcal{K}v_h, w_h)_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})v_h, w_h)_{L(\partial D)} + s_h(v_h, w_h)\big)$. The definition of r_h implies that $\Re(a_h^{\mathrm{Fl}}(v_h, w_h)) \le \|w_h\|_{V_h} r_h$, and (58.13) gives $|\Re(a_h^{\mathrm{Fl}}(v_h, w_h))| \le c c_2^{-\frac{1}{2}} l_h r_h$. Let us now estimate $|\Upsilon_h|$. Using the Cauchy–Schwarz inequality, the assumptions (57.33b)-(57.33c) for the boundary fields, and $\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})} \le c\mu_0$, we infer that

$$|\Upsilon_h| \le c\,\Big(\mu_0\|v_h\|_L\|w_h\|_L + |v_h|_{\mathcal{M}^{\mathrm{BP}}}\|\rho^{\frac{1}{2}}w_h\|_{L(\partial D)}\Big) + |v_h|_{\mathcal{S}}|w_h|_{\mathcal{S}},$$

with the scaling factor $\rho \in L^\infty(\partial D)$ s.t. $\rho_{|F} := \beta_{K_l}$ for all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. Using a discrete trace inequality shows that $\|\rho^{\frac{1}{2}}w_h\|_{L(\partial D)} \le c\|\tau^{-\frac{1}{2}}w_h\|_L$. Invoking (58.12) and (58.13) gives

$$\begin{aligned}|\Upsilon_h| &\le c\,\big(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |v_h|_{\mathcal{S}}^2\big)^{\frac{1}{2}}\Big(\mu_0\|w_h\|_L^2 + \|\tau^{-\frac{1}{2}}w_h\|_L^2 + |w_h|_{\mathcal{S}}^2\Big)^{\frac{1}{2}}\\ &\le c\,r_h^{\frac{1}{2}}l_h^{\frac{1}{2}}\Big(\|w_h\|_{V_h}^2 + \|\tau^{-\frac{1}{2}}w_h\|_L^2 + |w_h|_{\mathcal{S}}^2\Big)^{\frac{1}{2}} \le c\,r_h^{\frac{1}{2}}l_h^{\frac{3}{2}}.\end{aligned}$$

Thus, $\big|\Re\big((A_1(v_h), w_h)_L\big)\big| \le c(r_h l_h + r_h^{\frac{1}{2}} l_h^{\frac{3}{2}})$. Summing (58.12) and (58.14) yields $l_h^2 \le c(r_h l_h + r_h^{\frac{1}{2}} l_h^{\frac{3}{2}})$. We obtain the expected bound by applying Young's inequality twice. □

Remark 58.3 (Hermitian symmetry). The coercivity argument invoked in (58.12) shows that it is natural to assume that the sesquilinear form s_h is Hermitian symmetric since it is the real part of a_h^{Fl} that is L-coercive. □

58.3 Continuous interior penalty

The key idea in *CIP stabilization* (also termed *edge stabilization* in the literature) is to penalize the jump of $A_1(v_h)$ across the mesh interfaces. This idea has been introduced in Burman [57], Burman and Hansbo [67]. See also Burman and Ern [62, 63] for the *hp*-analysis and extensions to Friedrichs' systems, and [121] for nonlinear conservation laws.

58.3.1 Design of the CIP stabilization

Our goal is to construct a stabilization bilinear form s_h and an operator $\mathcal{J}_h$ that satisfy the conditions (58.4). We consider the discrete space

$$V_h := P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m), \tag{58.15}$$

where $P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m) := P_k^{\mathrm{g,b}}(\mathcal{T}_h;\mathbb{C}^m) \cap H^1(D;\mathbb{C}^m)$ and $P_k^{\mathrm{g,b}}(\mathcal{T}_h;\mathbb{C}^m)$ is the broken finite element space built using a reference finite element of degree $k \ge 1$ and the mesh $\mathcal{T}_h$ (see §19.2.1).

The main tool in the analysis of CIP is the discrete averaging operator

$$\mathcal{J}_h^{\mathrm{g,av}} : P_k^{\mathrm{g,b}}(\mathcal{T}_h;\mathbb{C}^m) \to P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m), \tag{58.16}$$

which acts componentwise as the discrete averaging operator introduced in §22.2. The essential idea in CIP is to build the linear operator $\mathcal{J}_h : V_h \to V_h$ used in the abstract design conditions (58.4) by setting $\mathcal{J}_h(v_h) := \mathcal{J}_h^{\mathrm{g,av}}\big(\tau A_1(v_h)\big)$ for all $v_h \in V_h$. Thus, the operator $\mathcal{J}_h$ produces an averaged version, scaled by the weighting parameter τ, of the gradient part of the differential operator in the Friedrichs' system.

The above definition though meets with two technical difficulties. The first one is that $A_1(v_h)$ is not a piecewise polynomial whenever the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are space-dependent. This can be fixed by setting

$$(\underline{A}_1(v_h))_{|K} \coloneqq \sum_{k\in\{1:d\}} \underline{\mathcal{A}}_K^k \partial_k v_{h|K}, \qquad \underline{\mathcal{A}}_K^k \coloneqq \frac{1}{|K|}\int_K \mathcal{A}^k \,\mathrm{d}x, \tag{58.17}$$

for all $v_h \in V_h$ and all $K \in \mathcal{T}_h$, and by replacing $A_1(v_h)$ by $\underline{A}_1(v_h)$. The second difficulty is that the local weighting parameter τ is by definition a discontinuous function across the mesh interfaces. To avoid dealing with the jumps of τ, we define the continuous mesh-dependent weighting function

$$\phi \coloneqq \mathcal{J}_{h,1}^{\mathrm{g,av}}(\tau) \in P_1^{\mathrm{g}}(\mathcal{T}_h), \tag{58.18}$$

where we use the scalar-valued discrete averaging operator of degree one. We have $\phi(\boldsymbol{z}) = \mathrm{card}(\mathcal{T}_{\boldsymbol{z}})^{-1}\sum_{K\in\mathcal{T}_{\boldsymbol{z}}}\tau_K$ with $\mathcal{T}_{\boldsymbol{z}} \coloneqq \{K \in \mathcal{T}_h \mid \boldsymbol{z} \in K\}$ for every mesh vertex $\boldsymbol{z}$. Recall the notation $\check{\mathcal{T}}_K \coloneqq \{K' \in \mathcal{T}_h \mid K \cap K' \neq \emptyset\}$ and $D_K \coloneqq \mathrm{int}(\bigcup_{K'\in\check{\mathcal{T}}_K} K')$ for all $K \in \mathcal{T}_h$. Notice that D_K represents a local neighborhood of K in D. We will also consider the slightly larger neighborhood $D_K^{(2)} \coloneqq \mathrm{int}(\bigcup_{K'\in\check{\mathcal{T}}_K^{(2)}} K')$ with $\check{\mathcal{T}}_K^{(2)} \coloneqq \{K' \in \mathcal{T}_h \mid \overline{D}_K \cap K' \neq \emptyset\}$. To avoid technicalities, we are going to assume that the piecewise constant function β_K is mildly graded. More precisely, we assume that there is c such that for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,

$$\beta_K \le c \min_{K'\in\check{\mathcal{T}}_K^{(2)}} \beta_{K'}. \tag{58.19}$$

The more general situation, which includes problems with contrasted coefficients, is further discussed in Remark 58.9.

Lemma 58.4 (Local bounds). *Let $\tau \in P_0^{\mathrm{b}}(\mathcal{T}_h)$ be defined in* (58.2) *and let $\phi \in P_1^{\mathrm{g}}(\mathcal{T}_h)$ be defined in* (58.18)*. Assume* (58.19)*. There is c s.t. the following holds true for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$:*

$$\|\phi\|_{L^\infty(D_K)} \le c \inf_{K'\in\check{\mathcal{T}}_K} \tau_{K'}, \tag{58.20a}$$

$$\|\phi^{-1}\|_{L^\infty(K)} \le c\,\tau_K^{-1}. \tag{58.20b}$$

Proof. See Exercise 58.2. □

We define $L(K) \coloneqq L^2(K;\mathbb{C}^m)$ and set $\|v\|_{L(K)} \coloneqq \|v\|_{L^2(K;\mathbb{C}^m)}$ for all $K \in \mathcal{T}_h$, and use a similar notation for $L(D_K)$ and $L(F)$ for all $F \in \mathcal{F}_h$. Recall that $[\![\cdot]\!]_F$ denotes the jump across the mesh interface $F \in \mathcal{F}_h^\circ$ (see Definition 18.2).

Proposition 58.5 ($\mathcal{J}_h$ for CIP). *Assume* (58.19) *and that s_h is defined so that there is c such that the following holds true for all $v_h \in V_h$ and all $h \in \mathcal{H}$:*

$$\sum_{F\in\mathcal{F}_h^\circ} \tau_F h_F \|[\![\underline{A}_1(v_h)]\!]_F\|_{L(F)}^2 \le c\left(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2\right), \tag{58.21}$$

with

$$\tau_F \coloneqq \max(\tau_{K_l}, \tau_{K_r}), \qquad \forall F \coloneqq \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ. \tag{58.22}$$

Then the conditions (58.4b)-(58.4c) *on s_h are satisfied with*

$$\mathcal{J}_h(v_h) \coloneqq \mathcal{J}_h^{\mathrm{g,av}}\big(\phi\underline{A}_1(v_h)\big). \tag{58.23}$$

Proof. (1) Let us prove (58.4b). Let $K \in \mathcal{T}_h$. We observe that the local L^2-stability of $\mathcal{J}_h^{\mathrm{g,av}}$ (see Corollary 22.4) implies that

$$\begin{aligned}\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_{L(K)}^2 &= \tau_K^{-1}\|\mathcal{J}_h^{\mathrm{g,av}}(\phi\underline{A}_1(v_h))\|_{L(K)}^2\\ &\le c\,\tau_K^{-1}\|\phi\underline{A}_1(v_h)\|_{L(D_K)}^2 \le c'\|\tau^{\frac12}\underline{A}_1(v_h)\|_{L(D_K)}^2,\end{aligned}$$

where we used (58.20a) in the last bound. Summing over $K \in \mathcal{T}_h$ and invoking the regularity of the mesh sequence, we infer that $\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L \le c\|\tau^{\frac12}\underline{A}_1(v_h)\|_L$. Using the triangle inequality then yields

$$\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L \le c\,(\|\tau^{\frac12}A_1(v_h)\|_L + \|\tau^{\frac12}(A_1-\underline{A}_1)(v_h)\|_L).$$

Since the fields $\mathcal{A}^k$ are piecewise Lipschitz with constant $L_{\mathcal{A}} \le c\mu_0$, using an inverse inequality and $\tau_K \le \mu_0^{-1}$ gives $\|\tau^{\frac12}(A_1-\underline{A}_1)(v_h)\|_L \le c\mu_0^{\frac12}\|v_h\|_L$. This proves (58.4b).
(2) Let us now prove (58.4c). Since (58.20b) implies that $\tau_K \le c\inf_{\boldsymbol{x}\in K}|\phi(\boldsymbol{x})|$, it is sufficient to bound $\|\phi^{\frac12}A_1(v_h)\|_L$. We first observe that

$$\|\phi^{\frac12}A_1(v_h)\|_L^2 = \mathfrak{T}_1 + \Re\Big(\big(A_1(v_h),\phi\underline{A}_1(v_h)-\mathcal{J}_h(v_h)\big)_L + \big(A_1(v_h),\phi(A_1-\underline{A}_1)(v_h)\big)_L\Big),$$

with $\mathfrak{T}_1 := \Re\big((A_1(v_h),\mathcal{J}_h(v_h))_L\big)$. Using Young's inequality gives

$$\frac12\|\phi^{\frac12}A_1(v_h)\|_L^2 \le \mathfrak{T}_1 + \|\phi^{-\frac12}(\phi\underline{A}_1(v_h)-\mathcal{J}_h(v_h))\|_L^2 + \|\phi^{\frac12}(A_1-\underline{A}_1)(v_h)\|_L^2.$$

Let us denote by $\mathfrak{T}_2,\mathfrak{T}_3$ the two rightmost terms on the right-hand side. Owing to Lemma 22.3 applied to the piecewise polynomial $\phi\underline{A}_1(v_h)$ and since ϕ is a continuous function, we infer that

$$\begin{aligned}\mathfrak{T}_2 &= \sum_{K\in\mathcal{T}_h}\|\phi^{-\frac12}(\phi\underline{A}_1(v_h)-\mathcal{J}_h^{\mathrm{g,av}}(\phi\underline{A}_1(v_h)))\|_{L(K)}^2\\ &\le c\sum_{K\in\mathcal{T}_h}\|\phi^{-1}\|_{L^\infty(K)}\sum_{F\in\check{\mathcal{F}}_K^\circ}\|\phi\|_{L^\infty(F)}^2 h_F\|[\![\underline{A}_1(v_h)]\!]_F\|_{L(F)}^2,\end{aligned}$$

where $\check{\mathcal{F}}_K^\circ := \{F\in\mathcal{F}_h^\circ \mid F\cap K\neq\emptyset\}$. Invoking the bound (58.20b) and since $\max_{F\in\check{\mathcal{F}}_K^\circ}\|\phi\|_{L^\infty(F)} \le \|\phi\|_{L^\infty(D_K)} \le c\tau_K$ by (58.20a), we infer that

$$\begin{aligned}\mathfrak{T}_2 &\le c\sum_{K\in\mathcal{T}_h}\tau_K\sum_{F\in\check{\mathcal{F}}_K^\circ}h_F\|[\![\underline{A}_1(v_h)]\!]_F\|_{L(F)}^2\\ &\le c\sum_{F\in\mathcal{F}_h^\circ}\tau_F h_F\|[\![\underline{A}_1(v_h)]\!]_F\|_{L(F)}^2 \le c\,\big(\mu_0\|v_h\|_L^2+|v_h|_{\mathcal{S}}^2\big),\end{aligned}$$

where we used the definition (58.22) of τ_F and the assumption (58.21). Finally, reasoning as above to estimate $(A_1-\underline{A}_1)(v_h)$ we obtain

$$\mathfrak{T}_3 \le c\sum_{K\in\mathcal{T}_h}\tau_K\mu_0^2\|v_h\|_{L(K)}^2 \le c\,\mu_0\|v_h\|_L^2.$$

Collecting the above bounds yields $\|\phi^{\frac12}A_1(v_h)\|_L^2 \le c(\mathfrak{T}_1+\mu_0\|v_h\|_L^2+|v_h|_{\mathcal{S}}^2)$. We conclude that (58.4c) holds true. □

Lemma 58.6 (s_h for CIP). *Assume* (58.19). *Let τ_F be defined in* (58.22). *The following sesquilinear form s_h^{CIP} satisfies* (58.4)*:*

$$s_h^{\mathrm{CIP}}(v_h,w_h) := \sum_{F\in\mathcal{F}_h^\circ} \tau_F h_F([\![\underline{A}_1(v_h)]\!]_F,[\![\underline{A}]\!]_1(w_h)_F)_{L(F)}. \tag{58.24}$$

Proof. (1) Condition (58.4a). Using that $\tau_F \le c\min(\tau_{K_l},\tau_{K_r})$, the triangle inequality to bound the jump, a discrete trace inequality, an inverse inequality, and the inequality $\beta_K h_K^{-1} \le \tau_K^{-1}$ which follows from (58.2), we infer that

$$\begin{aligned}|v_h|_{\mathcal{S}}^2 &\le \sum_{K\in\mathcal{T}_h} 2\tau_K h_K \|\underline{A}_1(v_h)\|_{L(\partial K)}^2 \le c\sum_{K\in\mathcal{T}_h} \tau_K \|\underline{A}_1(v_h)\|_{L(K)}^2\\ &\le c'\sum_{K\in\mathcal{T}_h} \tau_K\beta_K^2 h_K^{-2}\|v_h\|_{L(K)}^2 \le c'\|\tau^{-\frac12}v_h\|_L^2,\end{aligned}$$

for all $v_h\in V_h$. This proves (58.4a).
(2) The conditions (58.4b)-(58.4c) follow from Proposition 58.5 since (58.21) holds true with the definition (58.24). □

Remark 58.7 (Other example). It is also possible to consider the jumps of $A_1(v_h)$ in (58.24). Then (58.21) is shown by invoking as above that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are piecewise Lipschitz, $L_{\mathcal{A}}\le c\mu_0$, and $\mu_0\le\tau_K^{-1}$ for all $K\in\mathcal{T}_h$. Moreover, assuming that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are continuous over $\overline{D}$, another possibility is to set

$$s_h^{\mathrm{CIP}}(v_h,w_h) := \sum_{F\in\mathcal{F}_h^\circ} \beta_F h_F^2([\![\nabla v_h]\!]_F,[\![\nabla w_h]\!]_F)_{\boldsymbol{L}(F)}, \tag{58.25}$$

where $\beta_F := \max(\beta_{K_l},\beta_{K_r})$ with $F := \partial K_l\cap\partial K_r$; see Exercise 58.3. This choice is interesting for time-dependent fields $\mathcal{A}^k$ since the local assembling can be done only once, which is not the case for (58.24). □

Remark 58.8 (Simplified setting). Recall the simplified setting of Remark 58.1. Assume that sesquilinear form s_h is defined in (58.24) or (58.25). Let $\underline{A}_1(v_h)$ be defined in Proposition 58.5 for all $v_h\in V_h$. Then the operator $\mathcal{A}_h : V_h\to V_h$ s.t. $\mathcal{A}_h(v_h) := \mathcal{J}_h^{\mathrm{g,av}}(\underline{A}_1(v_h))\in V_h$ satisfies (58.7). □

Remark 58.9 (Contrasted coefficients). When solving problems with heterogeneous materials, the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$, and thus the coefficients $\{\beta_K\}_{K\in\mathcal{T}_h}$, can be strongly contrasted. In this case, the CIP stabilization can be designed using the above ideas provided the mesh cells can be organized into macroelements where the material properties are smooth. We refer the reader to Burman and Schieweck [69] for the analysis of CIP stabilization on composite elements. □

58.3.2 Error analysis

We assume in this section that s_h is defined in (58.24) or (58.25) and that (58.19) holds true. We perform the error analysis using Lemma 27.8. Since we have already established stability, it remains to bound the consistency error and prove convergence by using the approximation properties of finite elements (i.e., (58.1)). We assume that the solution to the model problem (57.1) has the following smoothness:

$$u\in V_{\mathrm{s}} := H^2(D;\mathbb{C}^m)\cap V. \tag{58.26}$$

We equip the space $V_\sharp := V_\mathrm{s} + V_h$ with the following two norms:

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac{1}{2}|v|_{\mathcal{M}^{\mathrm{BP}}}^2 + \|\tau^{\frac{1}{2}}A_1(v)\|_L^2, \tag{58.27a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \|v\|_{\tau,2}^2, \tag{58.27b}$$

with $\|v\|_{\tau,2}^2 := \sum_{K\in\mathcal{T}_h}\sum_{n\in\{0:2\}}\tau_K^{-1}h_K^{2n}|v|_{H^n(K)}^2$. Notice that (27.7) is satisfied with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ on $V_\sharp$). Notice also that the restriction of $\|\cdot\|_{V_\flat}$ to V_h is not $\|\cdot\|_{V_h}$, because we have dropped the seminorm $|\cdot|_{\mathcal{S}}$ in the definition of $\|\cdot\|_{V_\flat}$ (the reason for this is that $|\cdot|_{\mathcal{S}}$ may not be meaningful on V).

Lemma 58.10 (Consistency/boundedness). *Define the consistency error as*

$$\langle\delta_h(v_h), w_h\rangle_{V_h',V_h} := \ell_h^{\mathrm{Fl}}(w_h) - a_h^{\mathrm{Fl}}(v_h,w_h), \qquad \forall v_h, w_h \in V_h.$$

Assume that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are continuous over $\overline{D}$. There is $\omega_\sharp$, uniform w.r.t. $u \in V_\mathrm{s}$, such that for all $v_h, w_h \in V_h$ and all $h \in \mathcal{H}$,

$$|\langle\delta_h(v_h), w_h\rangle_{V_h',V_h}| \le \omega_\sharp\,\|u - v_h\|_{V_\sharp}\|w_h\|_{V_h}. \tag{58.28}$$

Proof. Since $A(u) = f$ in L and $(\mathcal{M}_F^{\mathrm{BP}} - \mathcal{N}_F)u = 0$ for all $F \in \mathcal{F}_h^\partial$ since $u \in V_\mathrm{s} \subset H^s(D;\mathbb{C}^m)$, $s > \frac{1}{2}$, we have

$$\langle\delta_h(v_h), w_h\rangle_{V_h',V_h} = (A(\eta), w_h)_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})\eta, w_h)_{L(\partial D)} - s_h(v_h,w_h),$$

with $\eta := u - v_h$. Using integration by parts, the first two terms on the right-hand side can be bounded as in the proof of Lemma 57.21 since $\|\tau^{-\frac{1}{2}}\eta\|_L + \|\rho^{\frac{1}{2}}\eta\|_{L(\partial D)} \le c\|\eta\|_{\tau,2}$ owing to the multiplicative trace inequality from Lemma 12.15 (with $p := 2$). Finally, we have $|s_h(v_h,w_h)| \le |v_h|_{\mathcal{S}}|w_h|_{\mathcal{S}}$. To bound the third factor, we observe that $|v_h|_{\mathcal{S}} = |\eta|_{\mathcal{S}}$ since $[\![\nabla u]\!]_F = 0$ because $u \in V_\mathrm{s}$, and Lemma 12.15 implies that $|\eta|_{\mathcal{S}} \le c\|\eta\|_{\tau,2}$. □

Theorem 58.11 (Error estimate). *Let the assumptions of Lemma* 58.10 *hold true.* (i) *There is c such that for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\flat} \le c\inf_{v_h\in V_h}\|u - v_h\|_{V_\sharp}. \tag{58.29}$$

(ii) *If $u \in H^{1+r}(D;\mathbb{C}^m)$, $r \in [1,k]$, then*

$$\|u - u_h\|_{V_\flat} \le c\left(\sum_{K\in\mathcal{T}_h}\max(\beta_K, \mu_0 h_K)h_K^{2r+1}|u|_{H^{1+r}(D_K;\mathbb{C}^m)}^2\right)^{\frac{1}{2}}. \tag{58.30}$$

Proof. The error estimate (58.29) follows from Lemma 27.8 and the above stability and consistency/boundedness results. The estimate (58.30) is obtained by using the approximation property (58.1) of the quasi-interpolation operator $\mathcal{I}_h$ and by proceeding as in the proof of Theorem 57.8 to estimate the various terms composing the $\|\cdot\|_{V_\sharp}$-norm. Finally, we use that $\tau_K^{-1}h_K = \max(\beta_K, \mu_0 h_K)$. □

Remark 58.12 (Simplified setting). The sesquilinear forms s_h defined in (58.24) or (58.25) satisfy the decay rates (58.8) in the simplified setting of Remark 58.1. Let us prove this claim. (i) We have $s_h(v,w_h) = 0$ for all $v \in H^{k+1}(D;\mathbb{C}^m)$ (recall that $k \ge 1$). Hence, $s_h(\mathcal{P}_{V_h}(v), w_h) =$

$s_h(\mathcal{P}_{V_h}(v)-v,w_h)$, and the estimate (58.8a) follows from the Cauchy–Schwarz inequality and the approximation properties of $\mathcal{P}_{V_h}$. (ii) With the operator $\mathcal{A}_h := \mathcal{J}_h^{\mathrm{g,av}}(\underline{A}_1) : V_h \to V_h$ from Remark 58.8 and since $h \le \ell_D$, we have

$$\|A_1(w_h)-\mathcal{A}_h(w_h)\|_L \le c\,\big(\tfrac{\beta}{h}\big)^{\frac12}\Big(|w_h|_{\mathcal{S}}+\big(\tfrac{\beta}{\ell_D}\big)^{\frac12}\|w_h\|_L\Big).$$

Hence, $(v-\mathcal{P}_{V_h}(v),A_1(w_h))_L = (v-\mathcal{P}_{V_h}(v),A_1(w_h)-\mathcal{A}_h(w_h))_L$, and (58.8b) follows from the Cauchy–Schwarz inequality. □

58.4 Examples

Example 58.13 (Advection-reaction). Consider the PDE $\mu u+\boldsymbol{\beta}\cdot\nabla u = f$ with the inflow boundary condition $u=0$ on ∂D^-; see §56.2.1. Assume that all the boundary faces of the mesh are subsets of either ∂D^- or $\partial D\backslash\partial D^-$. Let $\mu_0 := \operatorname{ess\,inf}_D(\mu-\frac12\nabla\cdot\boldsymbol{\beta})$, $\beta := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$, $\tau_K := (\beta_K^{-1}h_K,\mu_0^{-1})$ with $\beta_K := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$ for all $K\in\mathcal{T}_h$, and $\tau_F := \max(\tau_{K_l},\tau_{K_r})$ for all $F := \partial K_l\cap\partial K_r \in \mathcal{F}_h^\circ$. An example of stabilization bilinear form is

$$s_h^{\mathrm{CIP}}(v_h,w_h) := \sum_{F\in\mathcal{F}_h^\circ}\tau_F h_F(\llbracket\boldsymbol{\beta}\cdot\nabla v_h\rrbracket_F,\llbracket\boldsymbol{\beta}\cdot\nabla w_h\rrbracket_F)_{L^2(F)}.$$

The estimate of Theorem 58.11 with $r:=k$ gives

$$\mu_0^{\frac12}\|u-u_h\|_{L^2(D)} \le c\,\max(\beta,\mu_0 h)^{\frac12}h^{k+\frac12}|u|_{H^{k+1}(D)}.$$ □

Example 58.14 (Darcy). Consider the PDEs $\mathbb{d}^{-1}\boldsymbol{\sigma}+\nabla p = \boldsymbol{0}$ and $\mu p+\nabla\cdot\boldsymbol{\sigma} = f$ with the boundary condition $p=0$; see §56.2.2. Recalling the scaling argument from §57.3.3 and proceeding as in Example 57.12, we introduce the scaling matrix defined in (57.25) with the two reference scales d_* and μ_* (e.g., $d_*:=\lambda_\sharp$, $\mu_*:=\mu_\flat$). The (nondimensional) L-coercivity constant is $\mu_0 := \min(\frac{\mu_\flat}{\mu_*},\frac{d_*}{\lambda_\sharp})$. Setting $\ell_* := (d_*/\mu_*)^{\frac12}$, (57.24) implies that $\beta_K := \ell_*$, and the local weighting parameters are $\tau_F := \min(\ell_*^{-1}h_F,\mu_0^{-1})$ for all $F\in\mathcal{F}_h^\circ$. An example of stabilization bilinear form is

$$s_h^{\mathrm{CIP}}(v_h,w_h) := \sum_{F\in\mathcal{F}_h^\circ}\tau_F h_F\Big(d_*(\llbracket\nabla p_h\rrbracket,\llbracket\nabla q_h\rrbracket)_{\boldsymbol{L}^2(F)}+\mu_*^{-1}(\llbracket\nabla\cdot\boldsymbol{\sigma}_h\rrbracket,\llbracket\nabla\cdot\boldsymbol{\tau}_h\rrbracket)_{L^2(F)}\Big).$$

The estimate of Theorem 58.11 with $r:=k$ gives

$$\begin{aligned}&\mu_0^{\frac12}d_*^{-\frac12}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}+\mu_0^{\frac12}\mu_*^{\frac12}\|p-p_h\|_{L^2(D)}\\&+\ell_* d_*^{-\frac12}\|\tau^{\frac12}\nabla\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_{L^2(D)}+\ell_*\mu_*^{\frac12}\|\tau^{\frac12}\nabla(p-p_h)\|_{\boldsymbol{L}^2(D)}\\&\le c\,\max(\ell_*,\mu_0 h)^{\frac12}h^{k+\frac12}\Big(d_*^{-\frac12}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+1}(D)}+\mu_*^{\frac12}|p|_{H^{k+1}(D)}\Big).\end{aligned}$$ □

Example 58.15 (Maxwell). Consider the PDEs $\sigma\boldsymbol{E}-\nabla\times\boldsymbol{H}=\boldsymbol{f}$ and $\mathrm{i}\omega\mu\boldsymbol{H}+\nabla\times\boldsymbol{E}=\boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n}=\boldsymbol{0}$. Recall the reference scales σ_* and $\tilde\mu_*$ for σ and $\omega\mu$, respectively, and the (nondimensional) L-coercivity constant $\mu_0 := \frac{1}{\sqrt2}\min(\frac{\sigma_\flat}{\sigma_*},\frac{\omega\mu_\flat}{\tilde\mu_*})$. Setting $\ell_* := (\sigma_*\tilde\mu_*)^{-\frac12}$,

(57.24) implies that $\beta_K := \ell_*$ and the local weighting parameters are $\tau_F := \min(\ell_*^{-1}h_F, \mu_0^{-1})$ for all $F \in \mathcal{F}_h^\circ$. An example of stabilization sesquilinear form is

$$s_h^{\mathrm{CIP}}(v_h, w_h) := \sum_{F\in\mathcal{F}_h^\circ} \tau_F h_F \Big(\widetilde{\mu}_*^{-1}([\![\nabla\times\boldsymbol{E}_h]\!], [\![\nabla\times\boldsymbol{e}_h]\!])_{\boldsymbol{L}^2(F)} + \sigma_*^{-1}([\![\nabla\times\boldsymbol{H}_h]\!], [\![\nabla\times\boldsymbol{b}_h]\!])_{\boldsymbol{L}^2(F)} \Big).$$

The estimate of Theorem 58.11 with $r := k$ gives

$$\begin{aligned}
&\mu_0^{\frac12}\big(\sigma_*^{\frac12}\|\boldsymbol{E}-\boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)} + \tilde\mu_*^{\frac12}\|\boldsymbol{H}-\boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)}\big)\\
&+ \ell_*\sigma_*^{\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{E}-\boldsymbol{E}_h)\|_{\boldsymbol{L}^2(D)} + \ell_*\tilde\mu_*^{\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{H}-\boldsymbol{H}_h)\|_{\boldsymbol{L}^2(D)}\\
&\le c\,\max(\ell_*,\mu_0 h)^{\frac12}h^{k+\frac12}\big(\sigma_*^{\frac12}|\boldsymbol{E}|_{\boldsymbol{H}^{k+1}(D)} + \tilde\mu_*^{\frac12}|\boldsymbol{H}|_{\boldsymbol{H}^{k+1}(D)}\big).
\end{aligned}$$

□

Exercises

Exercise 58.1 (Simplified setting). Consider the setting of Remark 58.1 and assume that (58.7) holds true. Let $\mathcal{J}_h(v_h) := \frac{h}{\beta}\mathcal{A}_h(v_h)$ for all $v_h \in V_h$. (i) Prove (58.4b). (ii) Prove (58.4c).

Exercise 58.2 (Local bounds for CIP). The goal of this exercise is to prove Lemma 58.4. (i) Let $c_1 \le c_1'$ be positive real numbers. Let a_1, a_2 be two positive real numbers such that $c_1a_1 \le a_2 \le c_1'a_1$. Verify that there are positive constants c_2, c_2', only depending on c_1 and c_1', such that $c_2\min(a_1,b) \le \min(a_2,b) \le c_2'\min(a_1,b)$ for any positive real number b. (*Hint*: distinguish the four possible cases.) (ii) Assume (58.19). Prove that there is c such that $\tau_K \le c\min_{K'\in\check{\mathcal{T}}_K^{(2)}}\tau_{K'}$ for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$. (*Hint*: use Step (i) and the regularity of the mesh sequence.) (iii) Prove (58.20). (*Hint*: use Step (ii), $\|\phi\|_{L^\infty(D_K)} \le \max_{L\in\check{\mathcal{T}}_K^{(2)}}\tau_L$, and $\|\phi^{-1}\|_{L^\infty(K)} \le \max_{K'\in\check{\mathcal{T}}_K}\tau_{K'}^{-1}$.)

Exercise 58.3 (Full gradient). Prove (58.21) for CIP with (58.25).

Exercise 58.4 (1D advection, CIP). Let $D := (0,1)$, $f \in L^\infty(D)$, and a nonuniform mesh $\mathcal{T}_h$ of D with nodes $\{x_i\}_{i\in\{0:I+1\}}$ and local cells $K_{i+\frac12} := [x_i, x_{i+1}]$ of size $h_{i+\frac12} := x_{i+1}-x_i$, $\forall i \in \{0{:}I\}$. Let $h_i := \frac12(h_{i-\frac12}+h_{i+\frac12})$, $\forall i \in \{1{:}I\}$, be the length scale associated with the interfaces. Let $V_h := \{v_h \in P_1^{\mathrm{g}}(\mathcal{T}_h) \mid v_h(0) = 0\}$. Let $\beta \ne 0$. Consider the problem $\beta\partial_x u = f$, $u(0) = 0$. (i) Write the CIP formulation for the problem using (58.25) and let $u_h \in V_h$ be the discrete solution. (ii) Show that the discrete problem has a unique solution. (iii) Let $u_h := \sum_{i\in\{1:I+1\}}\mathsf{U}_i\varphi_i$ and $\mathsf{U}_0 := 0$. Write the equation satisfied by $\mathsf{U}_{i-2},\dots,\mathsf{U}_{i+2}$, $\forall i \in \{2{:}I{-}1\}$. (iv) Simplify the equation by assuming that the mesh is uniform and interpret the result in terms of finite differences. (*Hint*: compare the CIP stabilization with the second-order finite difference approximation of $|\beta|h^3\partial_{xxxx}u$.) *Note*: the term $|\beta|h^3\partial_{xxxx}u$ is often called *hyperviscosity* in the literature.

Chapter 59

Fluctuation-based stabilization (II)

In this chapter, we continue the unified analysis of fluctuation-based stabilization techniques for Friedrichs' systems. We now focus on two closely related stabilization techniques known in the literature as *local projection stabilization* (LPS) and *subgrid viscosity* (SGV). The key idea is to introduce a two-scale decomposition of the discrete H^1-conforming finite element space which leads to the notions of resolved and fluctuating (or subgrid) scales. Both stabilization techniques rely on a least-squares penalty: LPS penalizes the fluctuation of the gradient and SGV penalizes the gradient of the fluctuation. As for the CIP technique studied in the previous chapter, we verify that the abstract design conditions (58.4) are met with LPS and SGV.

59.1 Two-scale decomposition

The starting point is a two-scale decomposition of the H^1-conforming finite element space V_h:

$$V_h = R_h + B_h, \tag{59.1}$$

where the sum may not be direct. The discrete space R_h is viewed as the space of the *resolved* scales, and B_h is viewed as the space of the *fluctuating* (or *subgrid*) scales. It is important to realize that the degrees of freedom attached to B_h only serve to achieve stability, and that the approximation error is controlled by the best-approximation error in the space of the resolved scales R_h. We assume the following local approximation property in R_h: There is a quasi-interpolation operator $\mathcal{I}_h^{\mathrm{R}} : V \to R_h$ and a constant c s.t. the following holds true for all $r \in [0,k]$, all $l \in \{0{:}1+\lfloor r \rfloor\}$, all $v \in H^{1+r}(D;\mathbb{C}^m)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$:

$$|v - \mathcal{I}_h^{\mathrm{R}}(v)|_{H^l(K;\mathbb{C}^m)} \le c\, h_K^{1+r-l} |v|_{H^{1+r}(D_K;\mathbb{C}^m)}, \tag{59.2}$$

where $D_K := \mathrm{int}(\bigcup_{K' \in \check{\mathcal{T}}_K} K')$ with $\check{\mathcal{T}}_K := \{K' \in \mathcal{T}_h \mid K \cap K' \neq \emptyset\}$ is a local neighborhood of K.

We assume that the space of the fluctuating scales can be localized in the form $B_h := \bigoplus_{K \in \mathcal{T}_h} B_K$, where the functions in B_K are supported in K (one may think of the members of B_K as bubble-type functions, as shown in the examples given below). Since $r_h \in R_h$ is a continuous, piecewise polynomial function, the components of its gradient $\partial_i r_h$, $i \in \{1{:}d\}$, belong to a broken finite

element space $G_h := \bigoplus_{K\in\mathcal{T}_h} G_K$, where functions in G_K are supported in K. We then consider the local L-orthogonal projections

$$\pi_K^{\mathrm{B}} : L(K) \to B_K, \qquad \pi_K^{\mathrm{G}} : L(K) \to G_K, \qquad \forall K \in \mathcal{T}_h, \tag{59.3}$$

and their global counterparts $\pi_h^{\mathrm{B}} : L \to B_h$ and $\pi_h^{\mathrm{G}} : L \to G_h$ defined by setting $\pi_{h|K}^{\mathrm{B}} := \pi_K^{\mathrm{B}}$ and $\pi_{h|K}^{\mathrm{G}} := \pi_K^{\mathrm{G}}$ for all $K \in \mathcal{T}_h$.

The key assumption linking the local gradient space G_K to the local fluctuation space B_K is the following inf-sup condition introduced in [146, 144, 148] (see also [226]): There is $\gamma > 0$ s.t. for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,

$$\inf_{g\in G_K} \sup_{b\in B_K} \frac{|(b,g)_{L(K)}|}{\|g\|_{L(K)}\|b\|_{L(K)}} \geq \gamma, \tag{59.4}$$

or equivalently

$$\gamma \|g\|_{L(K)} \leq \|\pi_K^{\mathrm{B}}(g)\|_{L(K)}, \qquad \forall g \in G_K. \tag{59.5}$$

We consider the same local weighting parameter as in the previous chapters:

$$\tau_K := \min(\beta_K^{-1} h_K, \mu_0^{-1}), \qquad \forall K \in \mathcal{T}_h, \tag{59.6}$$

and the piecewise constant function $\tau : D \to \mathbb{R}$ s.t. $\tau_{|K} := \tau_K$ for all $K \in \mathcal{T}_h$.

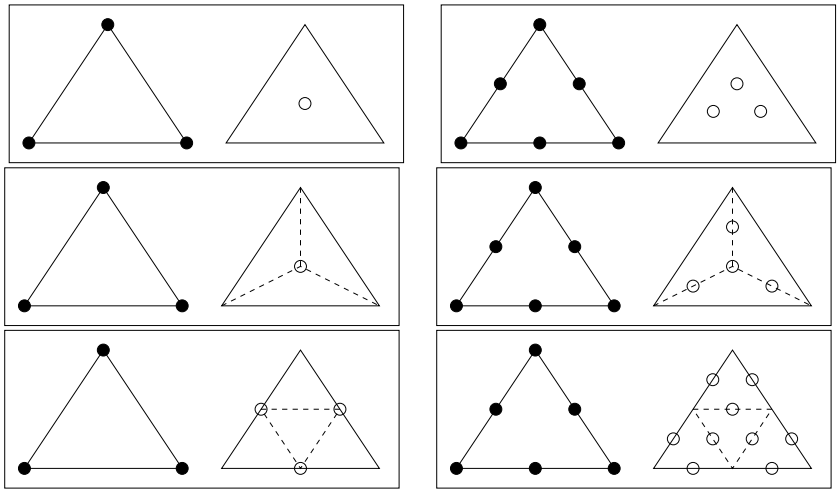

Figure 59.1: Two-scale finite elements. In each panel, the resolved scales are on the left and the fluctuating scales are on the right. The resolved scales are either $\mathbb{P}_1$ (left column) or $\mathbb{P}_2$ (right column) Lagrange elements. The upper panels illustrate the use of a standard bubble function to build the fluctuating scales. The central and the lower panels illustrate the use of piecewise polynomial bubble functions on a submesh with the same size (central panel) or half the size (bottom panel) to build the space of the resolved scales.

Let us describe three constructions of H^1-conforming finite element spaces of degree $k \geq 1$ which satisfy the above assumptions. (1) In the first example, the space of the resolved scales and that associated with the gradients are

$$R_h := P_k^{\mathrm{g}}(\mathcal{T}_h; \mathbb{C}^m), \qquad G_h := P_{k-1}^{\mathrm{b}}(\mathcal{T}_h; \mathbb{C}^m), \tag{59.7}$$

so that G_K is composed of $\mathbb{C}^m$-valued polynomials of degree at most $(k-1)$ on affine meshes. Following Guermond [144] for $k \in \{1,2\}$ and Matthies et al. [226] for all $k \geq 1$, we take $B_K := b_K G_K$, where b_K is the $H_0^1(K)$-bubble function proportional to the product of the $(d+1)$ barycentric coordinates over K as illustrated in the upper panels in Figure 59.1. (2) Instead of working with bubble functions, one can use hierarchical meshes; see [144, 226]. In this case, the construction starts from the mesh defining the space of the resolved scales, say $\check{\mathcal{T}}_h$. Assume for simplicity that $\check{\mathcal{T}}_h$ is composed of simplices. Then the mesh $\mathcal{T}_h$ defining V_h is built by barycentric refinement, i.e., for all $K \in \check{\mathcal{T}}_h$, $(d+1)$ new simplices are created by joining the barycenter of K to its $(d+1)$ vertices. Then we take

$$V_h := P_k^{\mathrm{g}}(\mathcal{T}_h;\mathbb{C}^m), \qquad R_h := P_k^{\mathrm{g}}(\check{\mathcal{T}}_h;\mathbb{C}^m), \qquad G_h := P_{k-1}^{\mathrm{b}}(\check{\mathcal{T}}_h;\mathbb{C}^m), \tag{59.8}$$

as shown in the panels in the second row of Figure 59.1. For all $K \in \check{\mathcal{T}}_h$, we have $\dim(P_k^{\mathrm{g}}(K;\mathbb{C}^m)) = m\binom{k+d}{d}$ and $g := \dim(G_K) = m\binom{k-1+d}{d}$; see (7.6). The number of shape functions in V_h that are supported in K is $g' := m(1+3\binom{k-1}{1}+3\binom{k-1}{2})$ in dimension 2 and $g' := m(1+4\binom{k-1}{1}+6\binom{k-1}{2}+4\binom{k-1}{3})$ in dimension 3. One always has $g' \geq g$. By working on the reference element, one can prove that among the g' shape functions that are supported in K one can always find g functions, say $\{\varphi_k^K\}_{k\in\{1:g\}}$, such that (59.4) holds true by setting $B_K := \mathrm{span}\{\varphi_1^K,\dots,\varphi_g^K\}$. The practical advantage of this construction is that V_h is a standard finite element space. (3) Finally, we mention the two-scale decomposition considered in [144] for $k \in \{1,2\}$ which also offers the advantage of V_h being a standard finite element space. A schematic representation of this decomposition is shown in the panels in the last row of Figure 59.1. The analysis (not considered here) is somewhat more involved since the fluctuating scales are represented by functions possibly supported in two adjacent mesh cells.

Remark 59.1 (Literature). The SGV technique has been introduced in Guermond [147, 146, 144, 148, 149] for monotone operators and semi-groups. The LPS technique has been introduced in Becker and Braack [28], Braack and Burman [41] for Stokes and convection-diffusion equations; see also Matthies et al. [226, 227]. The notion of scale separation and subgrid scale dissipation is similar in spirit to the *spectral viscosity* technique of Tadmor [269]. This notion is also found in the *Orthogonal Subscale Stabilization* technique of Codina [90]. □

59.2 Local projection stabilization

We define the fluctuation operator $\kappa_h^{\mathrm{G}} := I_L - \pi_h^{\mathrm{G}}$, where I_L is the identity operator in L.

Proposition 59.2 ($\mathcal{J}_h$ for LPS). *Assume that the inf-sup condition* (59.4) *is satisfied. Let τ_K be defined in* (59.6)*. Assume that the sesquilinear form s_h is defined so that there is $c>0$ s.t. for all $v_h \in V_h$ and all $h \in \mathcal{H}$,*

$$c\,\|\tau^{\frac12}\kappa_h^{\mathrm{G}}(A_1(v_h))\|_L^2 \leq \mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2. \tag{59.9}$$

Then the conditions (58.4b)-(58.4c) *are satisfied with the operator $\mathcal{J}_h : V_h \to V_h$ defined as follows:*

$$\mathcal{J}_h(v_h) := \tau\pi_h^{\mathrm{B}}\pi_h^{\mathrm{G}}(A_1(v_h)). \tag{59.10}$$

Proof. (1) We prove (58.4b) by using the local L-stability of π_h^{B} and π_h^{G}, i.e.,

$$\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L = \|\tau^{\frac12}\pi_h^{\mathrm{B}}\pi_h^{\mathrm{G}}(A_1(v_h))\|_L \leq \|\tau^{\frac12}A_1(v_h)\|_L. \tag{59.11}$$

(2) To prove (58.4c), we use the assumption (59.9) and the inf-sup condition (59.4) to infer that

$$\begin{aligned}\|\tau^{\frac12}A_1(v_h)\|_L^2 &= \|\tau^{\frac12}\pi_h^{\mathrm{G}}(A_1(v_h))\|_L^2 + \|\tau^{\frac12}\kappa_h^{\mathrm{G}}(A_1(v_h))\|_L^2\\ &\le \|\tau^{\frac12}\pi_h^{\mathrm{G}}(A_1(v_h))\|_L^2 + c\,(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal S}^2)\\ &\le \gamma^{-2}\|\tau^{\frac12}\pi_h^{\mathrm{B}}\pi_h^{\mathrm{G}}(A_1(v_h))\|_L^2 + c\,(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal S}^2).\end{aligned}$$

For the first term on the right-hand side, say $\mathfrak{T}_1$, we have

$$\begin{aligned}\gamma^2\mathfrak{T}_1 &= \Re\big((\pi_h^{\mathrm{B}}\pi_h^{\mathrm{G}}(A_1(v_h)),\mathcal J_h(v_h))_L\big) = \Re\big((\pi_h^{\mathrm{G}}(A_1(v_h)),\mathcal J_h(v_h))_L\big)\\ &= \Re\big((A_1(v_h),\mathcal J_h(v_h))_L\big) - \Re\big((\kappa_h^{\mathrm{G}}(A_1(v_h)),\mathcal J_h(v_h))_L\big),\end{aligned}$$

since $\mathcal J_h(v_h)\in B_h$. Let us set $\mathfrak{T}_2 := -\Re\big((\kappa_h^{\mathrm{G}}(A_1(v_h)),\mathcal J_h(v_h))_L\big)$. Then using the Cauchy–Schwarz inequality, Young's inequality, and (59.11), we obtain $|\mathfrak{T}_2|\le\delta\|\tau^{\frac12}A_1(v_h)\|_L^2 + c_\delta\|\tau^{\frac12}\kappa_h^{\mathrm{G}}(A_1(v_h))\|_L^2$ with $\delta>0$ as small as needed. The expected bound now follows from (59.9). □

As for CIP, it is convenient to filter out the local variations of the fields $\{\mathcal A^k\}_{k\in\{1:d\}}$ from the differential operator A_1. Thus, for all $v_h\in V_h$, we define $\underline{A}_1(v_h)$ as in (58.17) by setting for all $K\in\mathcal T_h$,

$$(\underline{A}_1(v_h))_{|K} := \sum_{k\in\{1:d\}}\underline{\mathcal A}_K^k\partial_k v_{h|K},\qquad \underline{\mathcal A}_K^k := \frac{1}{|K|}\int_K\mathcal A^k\,\mathrm{d}x. \tag{59.12}$$

Lemma 59.3 (s_h for LPS). *Let τ_K be defined in* (59.6)*. The following sesquilinear form s_h^{LPS} satisfies the conditions* (58.4)*:*

$$s_h^{\mathrm{LPS}}(v_h,w_h) := (\tau\kappa_h^{\mathrm{G}}(\underline{A}_1(v_h)),\kappa_h^{\mathrm{G}}(\underline{A}_1w_h))_L. \tag{59.13}$$

Proof. (1) The proof of (58.4a) follows from the L-stability of κ_h^{G}, an inverse inequality, and $\beta_K h_K^{-1}\le\tau_K^{-1}$.
(2) We now prove that (59.9) holds true and then invoke Proposition 59.2 to establish (58.4b)-(58.4c). The triangle inequality implies that

$$\|\tau^{\frac12}\kappa_h^{\mathrm{G}}(A_1(v_h))\|_L \le \|\tau^{\frac12}\kappa_h^{\mathrm{G}}(\underline{A}_1(v_h))\|_L + \|\tau^{\frac12}\kappa_h^{\mathrm{G}}((A_1-\underline{A}_1)(v_h))\|_L.$$

Using the L-stability of κ_h^{G}, the Lipschitz continuity of the fields $\mathcal A^k$, and the fact that $L_{\mathcal A}\le c\mu_0$ and $\tau_K\le\mu_0^{-1}$, we obtain $\|\tau^{\frac12}\kappa_h^{\mathrm{G}}(A_1(v_h))\|_L\le\|\tau^{\frac12}\kappa_h^{\mathrm{G}}(\underline{A}_1(v_h))\|_L + c\mu_0^{\frac12}\|v_h\|_L$. Hence, (59.9) holds true. □

Remark 59.4 (Other example). Notice that the choice (59.13) implies that $|r_h|_{\mathcal S}=0$ for all $r_h\in R_h$, i.e., $|\mathcal I_h^{\mathrm{R}}(u)|_{\mathcal S}=0$ for all $u\in V$. This property is important to establish the consistency of the approximation. Penalizing $\kappa_h^{\mathrm{G}}(A_1(v_h))$ instead of $\kappa_h^{\mathrm{G}}(\underline{A}_1(v_h))$ is somewhat delicate since $|\mathcal I_h^{\mathrm{R}}(u)|_{\mathcal S}$ no longer vanishes and bounding $|\mathcal I_h^{\mathrm{R}}(u)|_{\mathcal S}$ would require strong smoothness assumptions on the fields $\mathcal A^k$. Another possibility ensuring $|r_h|_{\mathcal S}=0$ for all $r_h\in R_h$ is to set

$$s_h^{\mathrm{LPS}}(v_h,w_h) := \sum_{K\in\mathcal T_h}\beta_K^2\tau_K(\kappa_h^{\mathrm{G}}(\nabla v_h),\kappa_h^{\mathrm{G}}(\nabla w_h))_{\boldsymbol L^2(K)}. \tag{59.14}$$

This choice is interesting for time-dependent fields $\mathcal A^k$ since the local assembling can be done only once, which is not the case for (59.13). □

Remark 59.5 (Simplified setting). Recall the setting of Remark 58.1. Let the sesquilinear form s_h be defined in (59.13) or (59.14). Then the operator $\mathcal A_h: V_h\to V_h$ s.t. $\mathcal A_h(v_h) := \pi_h^{\mathrm{G}}(\underline{A}_1(v_h))\in V_h$ satisfies (58.7). This follows from $\|A_1(v_h)-\mathcal A_h(v_h)\|_L\le\|A_1(v_h)-\underline{A}_1(v_h)\|_L+\|\kappa_h^{\mathrm{G}}(\underline{A}_1(v_h))\|_L$. □

59.3 Subgrid viscosity

In the *subgrid viscosity* method, the decomposition of V_h is assumed to be direct:

$$V_h = R_h \oplus B_h, \tag{59.15}$$

and to be locally L-stable, i.e., there is $\gamma_R > 0$ s.t. for all $v_h \in V_h$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,

$$\gamma_R \|\pi_h^{\mathrm{R}}(v_h)\|_{L(K)} \le \|v_h\|_{L(\tilde{D}_K)}, \tag{59.16}$$

where $\tilde{D}_K$ is a local neighborhood of K ($\tilde{D}_K := K$ for the four examples illustrated in the upper and middle panels of Figure 59.1, and $\tilde{D}_K := \{K' \in \mathcal{T}_h \mid K' \cap K \in \mathcal{F}_h^\circ\}$ for the other two examples shown in the lower panels). Letting $\pi_h^{\mathrm{R}} : V_h \to R_h$ be the oblique projection based on (59.15), we define the fluctuation operator $\kappa_h^{\mathrm{R}} := I_{V_h} - \pi_h^{\mathrm{R}}$, where I_{V_h} is the identity operator in V_h. Just as for LPS stabilization, we can choose $R_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$. Then $G_h = P_{k-1}^{\mathrm{b}}(\mathcal{T}_h)$, i.e., $G_K := \mathbb{P}_{k-1,d}$ on simplicial affine meshes. The simple choice $B_K := b_K G_K$ is only possible for $k \le d$, since otherwise the decomposition (59.15) is no longer direct. For $k \ge d+1$, a simple possibility to get around this technicality is to set $B_K := b_K^\alpha G_K$ with α equal to $\frac{k+1}{d+1}$ or to the smallest integer larger than $\frac{k}{d+1}$; see also Guermond [144, Prop. 4.1].

Proposition 59.6 ($\mathcal{J}_h$ for SGV). *Assume that the inf-sup condition* (59.4) *and the assumptions* (59.15)-(59.16) *hold true. Assume that s_h is defined so that there is $c > 0$ such that for all $v_h \in V_h$ and all $h \in \mathcal{H}$,*

$$c\,\|\tau^{\frac12} A_1(\kappa_h^{\mathrm{R}}(v_h))\|_L^2 \le \mu_0 \|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2. \tag{59.17}$$

Then the conditions (58.4b)-(58.4c) *are satisfied with the operator $\mathcal{J}_h : V_h \to V_h$ defined as follows:*

$$\mathcal{J}_h(v_h) := \tau \pi_h^{\mathrm{B}} \underline{A}_1(\pi_h^{\mathrm{R}}(v_h)). \tag{59.18}$$

Proof. (1) Proof of (58.4b). We have

$$\begin{aligned}
&\frac13 \|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L^2 \le \|\tau^{\frac12} A_1(v_h)\|_L^2 + \mathfrak{T}_1 + \mathfrak{T}_2,\\
&\mathfrak{T}_1 := \|\tau^{\frac12}(A_1 - \underline{A}_1)(\pi_h^{\mathrm{R}}(v_h))\|_L^2, \qquad \mathfrak{T}_2 := \|\tau^{\frac12} A_1(\kappa_h^{\mathrm{R}}(v_h))\|_L^2,
\end{aligned}$$

where we used the triangle inequality and the L-stability of π_h^{B}. The Lipschitz continuity of the fields $\mathcal{A}^k$, an inverse inequality, the L-stability of π_h^{R} from (59.16), and the inequalities $L_{\mathcal{A}} \le c\mu_0$ and $\tau_K \le \mu_0^{-1}$ imply $|\mathfrak{T}_1| \le c\mu_0 \|v_h\|_L^2$. The term $\mathfrak{T}_2$ is bounded by using the assumption (59.17). This proves (58.4b).

(2) Proof of (58.4c). Using the same definitions as above for $\mathfrak{T}_1$ and $\mathfrak{T}_2$, the triangle inequality yields $\frac13\|\tau^{\frac12} A_1(v_h)\|_L^2 \le \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3$ with $\mathfrak{T}_3 := \|\tau^{\frac12}\underline{A}_1(\pi_h^{\mathrm{R}}(v_h))\|_L^2$. Let us now estimate $\mathfrak{T}_3$. Since $\underline{A}_1(\pi_h^{\mathrm{R}}(v_h)) \in G_h$, we use the inf-sup condition (59.4) and the fact that $\mathcal{J}_h(v_h) \in B_h$ to infer that

$$\begin{aligned}
\gamma^2 \mathfrak{T}_3 &\le \|\tau^{\frac12}\pi_h^{\mathrm{B}}(\underline{A}_1(\pi_h^{\mathrm{R}}(v_h)))\|_L^2 = (\pi_h^{\mathrm{B}}(\underline{A}_1(\pi_h^{\mathrm{R}}(v_h))), \mathcal{J}_h(v_h))_L\\
&= (\underline{A}_1(\pi_h^{\mathrm{R}}(v_h)), \mathcal{J}_h(v_h))_L\\
&= \Re\big((A_1(\pi_h^{\mathrm{R}}(v_h)), \mathcal{J}_h(v_h))_L\big) - \Re\big(((A_1 - \underline{A}_1)(\pi_h^{\mathrm{R}}(v_h)), \mathcal{J}_h(v_h))_L\big)\\
&\le \Re\big((A_1(v_h), \mathcal{J}_h(v_h))_L\big) + \mathfrak{T}_4 + \mathfrak{T}_5,
\end{aligned}$$

with $\mathfrak{T}_4 := |(A_1(\kappa_h^{\mathrm{R}}(v_h)), \mathcal{J}_h(v_h))_L|$ and $\mathfrak{T}_5 := |((A_1 - \underline{A}_1)(\pi_h^{\mathrm{R}}(v_h)), \mathcal{J}_h(v_h))_L|$. We observe that $|\mathfrak{T}_4| \le \delta\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L^2 + c_\delta\|\tau^{\frac12} A_1(\kappa_h^{\mathrm{R}}(v_h))\|_L^2$ owing to the Cauchy–Schwarz and Young's inequalities, where $\delta > 0$ can be chosen as small as needed. Using the bound on $\|\tau^{-\frac12}\mathcal{J}_h(v_h)\|_L^2$ from

Step (1) together with (59.17), we infer that $|\mathfrak{T}_4| \le \delta\|\tau^{\frac12}A_1(v_h)\|_L^2 + c_\delta(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2)$. We proceed similarly for $\mathfrak{T}_5$ and use the above bound on $\mathfrak{T}_1$ to infer that $|\mathfrak{T}_5| \le \delta\|\tau^{\frac12}A_1(v_h)\|_L^2 + c_\delta\mu_0\|v_h\|_L^2$. Collecting these bounds leads to

$$\tfrac13\|\tau^{\frac12}A_1(v_h)\|_L^2 \le \delta\|\tau^{\frac12}A_1(v_h)\|_L^2 + \Re((A_1(v_h),\mathcal{J}_h(v_h))_L) + c_\delta(\mu_0\|v_h\|_L^2 + |v_h|_{\mathcal{S}}^2).$$

Choosing $\delta > 0$ sufficiently small leads to (58.4c). □

Lemma 59.7 (s_h for SGV). *Let τ_K be defined in* (59.6)*. The following sesquilinear form s_h^{SGV} satisfies the conditions* (58.4)*:*

$$s_h^{\mathrm{SGV}}(v_h,w_h) := (\tau\underline{A}_1(\kappa_h^{\mathrm{R}}(v_h)),\underline{A}_1(\kappa_h^{\mathrm{R}}(w_h)))_L. \tag{59.19}$$

Proof. See Exercise 59.3. □

Remark 59.8 (Other example). Another possibility is to set

$$s_h^{\mathrm{SGV}}(v_h,w_h) := \sum_{K\in\mathcal{T}_h}\beta_K^2\tau_K(\nabla(\kappa_h^{\mathrm{R}}(v_h)),\nabla(\kappa_h^{\mathrm{R}}(w_h)))_{\boldsymbol{L}(K)}. \tag{59.20}$$

This choice is interesting for time-dependent fields $\mathcal{A}^k$ since the local assembling can be done only once, which is not the case for (59.19). □

Remark 59.9 (Simplified setting). Recall the setting of Remark 58.1. Let the sesquilinear form s_h be defined in (59.19) or (59.20). Then the operator $\mathcal{A}_h : V_h \to V_h$ s.t. $\mathcal{A}_h(v_h) := \pi_h^{\mathrm{G}}(\underline{A}_1(v_h))$ satisfies (58.7). This operator is the same as for LPS, but the proof of (58.7) is slightly different. For SGV, we have $\mathcal{A}_h(v_h) = \underline{A}_1(\pi_h^{\mathrm{R}}(v_h)) + \pi_h^{\mathrm{G}}(\underline{A}_1(\kappa_h^{\mathrm{R}}(v_h)))$ since $\underline{A}_1(\pi_h^{\mathrm{R}}(v_h)) \in G_h$, so that $\underline{A}_1(v_h) - \mathcal{A}_h(v_h) = (I_L - \pi_h^{\mathrm{G}})(\underline{A}_1(\kappa_h^{\mathrm{R}}(v_h)))$. Hence, $\|\underline{A}_1(v_h) - \mathcal{A}_h(v_h)\|_L \le \|\underline{A}_1(\kappa_h^{\mathrm{R}}(v_h))\|_L \le c\left(\frac{\beta}{h}\right)^{\frac12}|v_h|_{\mathcal{S}}$. □

59.4 Error analysis

The error analysis proceeds as in the proof of Lemma 27.8 with one modification: we consider the best-approximation error of u in R_h and not in V_h. This choice is reasonable since the space of the resolved scales R_h has optimal approximation properties as assumed in (59.2). We assume that

$$u \in V_{\mathrm{S}} := H^s(D;\mathbb{C}^m)\cap V, \qquad s > \frac12. \tag{59.21}$$

We equip $V_\sharp := V_{\mathrm{S}} + V_h$ with the norms

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \tfrac12|v|_{\mathcal{M}^{\mathrm{BP}}}^2 + \|\tau^{\frac12}A_1(v)\|_L^2, \tag{59.22a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \|\tau^{-\frac12}v\|_L^2 + \|\rho^{\frac12}v\|_{L(\partial D)}^2. \tag{59.22b}$$

Notice that $\|\cdot\|_{V_\flat}$ is the same norm as for CIP (see (58.27a)), and $\|\cdot\|_{V_\sharp}$ is the same norm as for GaLS (see (57.40b)). Notice also that (27.7) is satisfied with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ on $V_\sharp$). Notice also that the restriction of $\|\cdot\|_{V_\flat}$ to V_h is not $\|\cdot\|_{V_h}$ since we have dropped the seminorm $|\cdot|_{\mathcal{S}}$ in the definition of $\|\cdot\|_{V_\flat}$ (the reason for this is that $|\cdot|_{\mathcal{S}}$ may not be meaningful on V).

Lemma 59.10 (Consistency/boundedness). *Assume that*

$$s_h(v_h, w_h) = 0, \qquad \forall (v_h, w_h) \in R_h \times V_h. \tag{59.23}$$

Define the consistency error as

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} := \ell_h^{\mathrm{Fl}}(w_h) - a_h^{\mathrm{Fl}}(v_h, w_h), \qquad \forall (v_h, w_h) \in R_h \times V_h.$$

There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{S}}$, such that for all $(v_h, w_h) \in R_h \times V_h$ and all $h \in \mathcal{H}$,

$$|\langle \delta_h(v_h), w_h \rangle_{V_h', V_h}| \le \omega_\sharp \, \|u - v_h\|_{V_\sharp} \|w_h\|_{V_h}. \tag{59.24}$$

Proof. Proceed as in the proof of Lemma 58.10 except that we now have $s_h(v_h, w_h) = 0$ for all $(v_h, w_h) \in R_h \times V_h$ by assumption. □

Theorem 59.11 (Error estimate). *Let the assumptions of Lemma* 59.10 *hold true.* (i) *There is c such that for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\flat} \le c \inf_{v_h \in R_h} \|u - v_h\|_{V_\sharp}. \tag{59.25}$$

(ii) *If $u \in H^{1+r}(D; \mathbb{C}^m)$, $r \in [0, k]$, then*

$$\|u - u_h\|_{V_\flat} \le c \Big(\sum_{K \in \mathcal{T}_h} (\tau_K^{-1} h_K) h_K^{2r+1} |u|^2_{H^{1+r}(D_K; \mathbb{C}^m)} \Big)^{\frac{1}{2}}. \tag{59.26}$$

Proof. Similar to that of Theorem 58.11 except that we now invoke the approximation properties (59.2) of the quasi-interpolation operator $\mathcal{I}_h^{\mathrm{R}}$. □

Remark 59.12 (Simplified setting). The sesquilinear forms s_h defined in (59.13) or (59.14) for LPS or in (59.19) or (59.20) for SGV satisfy the decay rates (58.8) in the simplified setting of Remark 58.1. Let us prove this claim. (i) We have $s_h(\mathcal{I}_h^{\mathrm{R}}(v), w_h) = 0$ for all $v \in H^1(D; \mathbb{C}^m)$. Hence, $s_h(\mathcal{P}_{V_h}(v), w_h) = s_h(\mathcal{P}_{V_h}(v) - v, w_h) + s_h(v - \mathcal{I}_h^{\mathrm{R}}(v), w_h)$, and (76.20b) follows from the Cauchy–Schwarz inequality and the approximation properties of $\mathcal{P}_{V_h}$ and $\mathcal{I}_h^{\mathrm{R}}$. (ii) With the operator $\mathcal{A}_h : V_h \to V_h$ defined in Remark 59.5 for LPS and in Remark 59.9 for SGV, and since $h \le \ell_D$, we have

$$\|A_1(w_h) - \mathcal{A}_h(w_h)\|_L \le c \, \big(\tfrac{\beta}{h}\big)^{\frac{1}{2}} \Big(|w_h|_{\mathcal{S}} + \big(\tfrac{\beta}{\ell_D}\big)^{\frac{1}{2}} \|w_h\|_L \Big).$$

We infer that $(v - \mathcal{P}_{V_h}(v), A_1(w_h))_L = (v - \mathcal{P}_{V_h}(v), A_1(w_h) - \mathcal{A}_h(w_h))_L$, and (58.8b) follows from the Cauchy–Schwarz inequality. □

59.5 Examples

Example 59.13 (Advection-reaction). Consider the PDE $\mu u + \boldsymbol{\beta} \cdot \nabla u = f$ with the inflow boundary condition $u = 0$ on ∂D^-; see §56.2.1. Assume that all the boundary faces of the mesh are subsets of either ∂D^- or $\partial D \backslash \partial D^-$. Let $\mu_0 := \operatorname{ess\,inf}_D (\mu - \frac{1}{2} \nabla \cdot \boldsymbol{\beta})$, $\beta := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$, and $\tau_K := (\beta_K^{-1} h, \mu_0^{-1})$ with $\beta_K := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$ for all $K \in \mathcal{T}_h$. Examples of stabilization bilinear forms are

$$s_h^{\mathrm{LPS}}(v_h, w_h) := \big(\tau \kappa_h^{\mathrm{G}}(\underline{\boldsymbol{\beta}} \cdot \nabla v_h), \kappa_h^{\mathrm{G}}(\underline{\boldsymbol{\beta}} \cdot \nabla w_h) \big)_{L^2},$$
$$s_h^{\mathrm{SGV}}(v_h, w_h) := \big(\tau \boldsymbol{\beta} \cdot \nabla(\kappa_h^{\mathrm{R}}(v_h)), \boldsymbol{\beta} \cdot \nabla(\kappa_h^{\mathrm{R}}(w_h)) \big)_{L^2}.$$

The estimate of Theorem 59.11 with $r := k$ gives

$$\mu_0^{\frac12}\|u-u_h\|_{L^2(D)} \le c\,\max(\beta,\mu_0 h)^{\frac12} h^{k+\frac12}|u|_{H^{k+1}(D)}.$$

This is the same estimate as with CIP; see Example 58.13. □

Example 59.14 (Darcy). Consider the PDEs $\mathbb{d}^{-1}\boldsymbol{\sigma}+\nabla p=\boldsymbol{0}$ and $\mu p+\nabla\cdot\boldsymbol{\sigma}=f$ with the boundary condition $p=0$; see §56.2.2. Recalling the scaling argument from §57.3.3 and proceeding as in Example 57.12, we introduce the scaling matrix defined in (57.25) with the two reference scales d_* and μ_* (e.g., $d_*:=\lambda_\sharp$, $\mu_*:=\mu_\flat$). The (nondimensional) L-coercivity constant is $\mu_0:=\min(\frac{\mu_\flat}{\mu_*},\frac{d_*}{\lambda_\sharp})$. Setting $\ell_*:=(d_*/\mu_*)^{\frac12}$, (57.24) implies that $\beta_K:=\ell_*$, and the local weighting parameter is $\tau_K:=\min(\ell_*^{-1}h_K,\mu_0^{-1})$ for all $K\in\mathcal{T}_h$. Examples of stabilization bilinear forms are

$$\begin{aligned}
s_h^{\mathrm{LPS}}(v_h,w_h)&:=d_*\big(\tau\kappa_h^{\mathrm{G}}(\nabla p_h),\kappa_h^{\mathrm{G}}(\nabla q_h)\big)_{\boldsymbol{L}^2}+\mu_*^{-1}\big(\tau\kappa_h^{\mathrm{G}}(\nabla\cdot\boldsymbol{\sigma}_h),\kappa_h^{\mathrm{G}}(\nabla\cdot\boldsymbol{\tau}_h)\big)_{L^2},\\
s_h^{\mathrm{SGV}}(v_h,w_h)&:=d_*\big(\tau\nabla(\kappa_h^{\mathrm{R}}(p_h)),\nabla(\kappa_h^{\mathrm{R}}(q_h))\big)_{\boldsymbol{L}^2}+\mu_*^{-1}\big(\tau\nabla\cdot(\kappa_h^{\mathrm{R}}(\boldsymbol{\sigma}_h)),\nabla\cdot(\kappa_h^{\mathrm{R}}(\boldsymbol{\tau}_h))\big)_{L^2}.
\end{aligned}$$

The estimate of Theorem 59.11 with $r:=k$ gives

$$\begin{aligned}
&\mu_0^{\frac12}d_*^{-\frac12}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}+\mu_0^{\frac12}\mu_*^{\frac12}\|p-p_h\|_{L^2(D)}\\
&+\ell_* d_*^{-\frac12}\|\tau^{\frac12}\nabla\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_{L^2(D)}+\ell_*\mu_*^{\frac12}\|\tau^{\frac12}\nabla(p-p_h)\|_{\boldsymbol{L}^2(D)}\\
&\le c\,\max(\ell_*,\mu_0 h)^{\frac12}h^{k+\frac12}\big(d_*^{-\frac12}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+1}(D)}+\mu_*^{\frac12}|p|_{H^{k+1}(D)}\big).
\end{aligned}$$

This is the same estimate as with CIP; see Example 58.14. □

Example 59.15 (Maxwell). Consider the PDEs $\sigma\boldsymbol{E}-\nabla\times\boldsymbol{H}=\boldsymbol{f}$ and $\mathrm{i}\omega\mu\boldsymbol{H}+\nabla\times\boldsymbol{E}=\boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n}=\boldsymbol{0}$. Recall the reference scales σ_* and $\tilde\mu_*$ for σ and $\omega\mu$, respectively, and the (nondimensional) L-coercivity constant $\mu_0:=\frac{1}{\sqrt2}\min(\frac{\sigma_\flat}{\sigma_*},\frac{\omega\mu_\flat}{\tilde\mu_*})$. Setting $\ell_*:=(\sigma_*\tilde\mu_*)^{-\frac12}$, (57.24) implies $\beta_K:=\ell_*$ so that $\tau_K:=\min(\ell_*^{-1}h_K,\mu_0^{-1})$ for all $K\in\mathcal{T}_h$. Examples of stabilization sesquilinear forms are

$$\begin{aligned}
s_h^{\mathrm{LPS}}(v_h,w_h)&:=\widetilde{\mu}_*^{-1}\big(\tau\kappa_h^{\mathrm{G}}(\nabla\times\boldsymbol{E}_h),\kappa_h^{\mathrm{G}}(\nabla\times\boldsymbol{e}_h)\big)_{\boldsymbol{L}^2}+\sigma_*^{-1}\big(\tau\kappa_h^{\mathrm{G}}(\nabla\times\boldsymbol{H}_h),\kappa_h^{\mathrm{G}}(\nabla\times\boldsymbol{b}_h)\big)_{\boldsymbol{L}^2},\\
s_h^{\mathrm{SGV}}(v_h,w_h)&:=\widetilde{\mu}_*^{-1}\big(\tau\nabla\times(\kappa_h^{\mathrm{R}}(\boldsymbol{E}_h)),\nabla\times(\kappa_h^{\mathrm{R}}(\boldsymbol{e}_h))\big)_{\boldsymbol{L}^2}+\sigma_*^{-1}\big(\tau\nabla\times(\kappa_h^{\mathrm{R}}(\boldsymbol{H}_h)),\nabla\times(\kappa_h^{\mathrm{R}}(\boldsymbol{b}_h))\big)_{\boldsymbol{L}^2}.
\end{aligned}$$

The estimate of Theorem 59.11 with $r:=k$ gives

$$\begin{aligned}
&\mu_0^{\frac12}\big(\sigma_*^{\frac12}\|\boldsymbol{E}-\boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)}+\tilde\mu_*^{\frac12}\|\boldsymbol{H}-\boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)}\big)\\
&+\ell_*\sigma_*^{\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{E}-\boldsymbol{E}_h)\|_{\boldsymbol{L}^2(D)}+\ell_*\tilde\mu_*^{\frac12}\|\tau^{\frac12}\nabla\times(\boldsymbol{H}-\boldsymbol{H}_h)\|_{\boldsymbol{L}^2(D)}\\
&\le c\,\max(\ell_*,\mu_0 h)^{\frac12}h^{k+\frac12}\big(\sigma_*^{\frac12}|\boldsymbol{E}|_{\boldsymbol{H}^{k+1}(D)}+\tilde\mu_*^{\frac12}|\boldsymbol{H}|_{\boldsymbol{H}^{k+1}(D)}\big).
\end{aligned}$$

This is the same estimate as with CIP; see Example 58.15. □

Exercises

Exercise 59.1 (Inf-sup condition). Consider the setting of §59.1 and assume that the functions in B_h vanish on ∂D. Prove that there is $\alpha>0$ such that for all $r_h\in R_h$ and all $h\in\mathcal{H}$,

$$\alpha(\|r_h\|_{V_h}+\mu_0^{-\frac12}\|A_1(r_h)\|_L)\le\sup_{w_h\in V_h}\frac{|a_h^{\mathrm{BP}}(r_h,w_h)|}{\|w_h\|_{V_h}},$$

with a_h^{BP} defined in (58.3) and $\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}}^2 + |v_h|_{\mathcal{S}^\partial}^2$ for all $v_h \in V_h$. (*Hint*: use the coercivity of a_h^{BP} to control $\|r_h\|_{V_h}$, and use that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are piecewise Lipschitz together with (59.4) to control $\mu_0^{-\frac{1}{2}}\|A_1(r_h)\|_L$.)

Exercise 59.2 (Full gradient). Prove (59.9) for the choice of s_h^{LPS} in Example 59.4.

Exercise 59.3 (SGV). Prove Lemma 59.7.

Chapter 60

Discontinuous Galerkin

In this chapter, we want to approximate the same model problem as in the previous two chapters, i.e., (57.1), but instead of using stabilized H^1-conforming finite elements we consider the discontinuous Galerkin (dG) method. The stability and convergence properties of the method rely on choosing a numerical flux across the mesh interfaces. Choosing the centered flux yields suboptimal convergence rates for smooth solutions. The stability properties of the method are tightened by penalizing the interface jumps, which corresponds to upwinding in the case of advection-reaction equations. The method thus obtained is called *upwind dG* irrespective of the nature of the PDE. This method gives the same error estimates as those obtained with stabilized H^1-conforming finite elements. Here again, the boundary conditions are enforced by the boundary penalty technique of §57.4.

60.1 Discrete setting

The dG method uses the broken finite element space

$$V_h := P_k^{\mathrm{b}}(\mathcal{T}_h;\mathbb{C}^m) := \{v_h \in L^\infty(D;\mathbb{C}^m) \mid v_{h|K} \in P_K,\, \forall K \in \mathcal{T}_h\}, \tag{60.1}$$

as discrete trial and test space. $P_k^{\mathrm{b}}(\mathcal{T}_h;\mathbb{C}^m)$ is built by using a finite element of degree $k \geq 0$ and a shape-regular sequence of affine meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly. (More general meshes can be considered as well.) The above assumptions imply that there is an interpolation operator $\mathcal{I}_h^{\mathrm{b}} : L \to V_h$ (one can consider the L-orthogonal projection onto V_h) s.t.

$$\|v - \mathcal{I}_h^{\mathrm{b}}(v)\|_{L(K)} + h_K\|\nabla(v - \mathcal{I}_h^{\mathrm{b}}(v))\|_{\boldsymbol{L}(K)} \leq c\, h_K^{1+r}|v|_{H^{1+r}(K;\mathbb{C}^m)}, \tag{60.2}$$

for all $r \in [0,k]$, all $v \in H^{1+r}(K;\mathbb{C}^m)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$.

The notions of jump (see Definition 18.2) and average (see Definition 38.1) play an important role in dG methods. We recall that $\mathcal{F}_h$ denotes the set of the mesh faces. This set is split into the subset of the mesh interfaces $\mathcal{F}_h^\circ$ and the subset of the boundary faces $\mathcal{F}_h^\partial$. Each mesh face is oriented by the fixed unit normal vector $\boldsymbol{n}_F$ with Cartesian components $(n_{F,k})_{k\in\{1:d\}}$. We define $L(F) := L^2(F;\mathbb{C}^m)$ for all $F \in \mathcal{F}_h$, $L(K) := L^2(K;\mathbb{C}^m)$ for all $K \in \mathcal{T}_h$, and we denote by $\mathcal{F}_K$ the collection of the faces of K. In the entire chapter, we assume that the fields $\mathcal{A}^k$ are smooth enough so that the Hermitian fields

$$\mathcal{N}_F := \sum_{k\in\{1:d\}} n_{F,k}\mathcal{A}^k_{|F}, \qquad \forall F \in \mathcal{F}_h^\circ, \tag{60.3}$$

are single-valued and are in $L^\infty(F;\mathbb{C}^{m\times m})$. We also write $\mathcal{N}_F := \mathcal{N}_{|F}$ for all $F \in \mathcal{F}_h^\partial$, with $\mathcal{N}$ defined in (56.3). Recall that

$$A(v) := \mathcal{K}v + A_1(v), \qquad A_1(v) := \sum_{k\in\{1:d\}} \mathcal{A}^k\partial_k v. \tag{60.4}$$

The formal adjoint $\tilde{A}$ of A is defined by $\tilde{A}(v) := (\mathcal{K}^\mathsf{H} - \mathcal{X})v - A_1(v)$ where $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k\mathcal{A}^k$. Similarly to the notion of broken gradient (see Definition 36.3), we define the broken differential operator $A_{1h} : V_h \to L$ s.t. $A_{1h}(v_h)_{|K} := A_1(v_{h|K})$ for all $v_h \in V_h$ and all $K \in \mathcal{T}_h$. We then set $A_h(v_h) := \mathcal{K}v_h + A_{1h}(v_h)$ and $\tilde{A}_h(v_h) := (\mathcal{K}^\mathsf{H} - \mathcal{X})v_h - A_{1h}(v_h)$. The following integration by parts formula (56.5) will be essential in this chapter.

Lemma 60.1 (Integration by parts). *Letting*

$$n_h(v_h, w_h) := \sum_{F\in\mathcal{F}_h^\circ} (\mathcal{N}_F[\![v_h]\!], \{w_h\})_{L(F)}, \tag{60.5}$$

for all $v_h, w_h \in V_h$, the following holds true:

$$\begin{aligned}(A_h(v_h), w_h)_L =& (v_h, \tilde{A}_h(w_h))_L + (\mathcal{N}v_h, w_h)_{L(\partial D)}\\ &+ n_h(v_h, w_h) + \overline{n_h(w_h, v_h)}.\end{aligned} \tag{60.6}$$

Proof. For all $K \in \mathcal{T}_h$, let $\boldsymbol{n}_K$ be the outward unit normal to K and set $\epsilon_{K,F} := \boldsymbol{n}_K\cdot\boldsymbol{n}_F = \pm 1$ for all $F \in \mathcal{F}_K$. Then $\epsilon_{K,F}\mathcal{N}_F = \mathcal{N}_K$, where $\mathcal{N}_K := \sum_{k\in\{1:d\}} n_{K,k}\mathcal{A}^k_{|K}$ and $(n_{K,k})_{k\in\{1:d\}}$ are the Cartesian components of $\boldsymbol{n}_K$. Proceeding as in the proof of Lemma 56.1, we infer that

$$(A(v_h), w_h)_{L(K)} = (v_h, \tilde{A}(w_h))_{L(K)} + \sum_{F\in\mathcal{F}_K} \epsilon_{K,F}(\mathcal{N}_F v_h, w_h)_{L(F)}.$$

We obtain (60.6) by summing this identity over the mesh cells and using the following properties for all $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$: $\epsilon_{K_l,F}(w_h^\mathsf{H}\mathcal{N}_F v_h)_{|K_l} + \epsilon_{K_r,F}(w_h^\mathsf{H}\mathcal{N}_F v_h)_{|K_r} = [\![w_h^\mathsf{H}\mathcal{N}_F v_h]\!]$ a.e. on F; $[\![w_h^\mathsf{H}\mathcal{N}_F v_h]\!] = \{w_h\}^\mathsf{H}\mathcal{N}_F[\![v_h]\!] + [\![w_h]\!]^\mathsf{H}\mathcal{N}_F\{v_h\}$ since $\mathcal{N}_F$ is single-valued by assumption; $[\![w_h]\!]^\mathsf{H}\mathcal{N}_F\{v_h\} = \overline{\{v_h\}^\mathsf{H}\mathcal{N}_F[\![w_h]\!]}$ since $\mathcal{N}_F$ is Hermitian. □

The sesquilinear form n_h can be extended to $V_\mathrm{s}\times V_\mathrm{s}$ with

$$V_\mathrm{s} := H^s(D;\mathbb{C}^m)\cap V, \qquad s > \frac{1}{2}. \tag{60.7}$$

In this case, $A_h(v) = A(v)$, $\tilde{A}_h(v) = \tilde{A}(v)$ for all $v \in V_\mathrm{s}$, and the integration by parts formula (60.6) reduces to (56.5) as we now show.

Corollary 60.2 (Jumps in graph space). *Let $v \in V_\mathrm{s}$. Then $\mathcal{N}_F[\![v]\!] = 0$ for all $F \in \mathcal{F}_h^\circ$.*

Proof. The proof is similar to that of Theorem 18.8. Let $\varphi \in C_0^\infty(D;\mathbb{C}^m)$. Applying (60.6) with $w := \varphi$ and using that since $[\![\varphi]\!]_F = 0$ for all $F \in \mathcal{F}_h^\circ$ and $\varphi_{|\partial D} = 0$ gives $(A(v),\varphi)_L = (v, \tilde{A}(\varphi))_L + \sum_{F\in\mathcal{F}_h^\circ}(\mathcal{N}_F[\![v]\!],\varphi)_{L(F)}$. But we also have $(A(v),\varphi)_L = (v,\tilde{A}(\varphi))_L$, whence the assertion. □

In the entire chapter, the boundary conditions are going to be enforced weakly by using the boundary penalty field $\mathcal{M}_F^\mathrm{BP} := \mathcal{M}_F + \mathcal{S}_F^\partial$ introduced in §57.4.2 and satisfying the assumptions stated in (57.33).

60.2 Centered fluxes

In this section, we study a dG method based on the use of centered fluxes.

60.2.1 Local and global formulation

Since it is possible to localize the functions in V_h to any cell $K \in \mathcal{T}_h$, a natural starting point of the dG method consists of looking for a local formulation. Let us assume that $u \in V_{\rm s}$ with $V_{\rm s}$ defined in (60.7). Let $K \in \mathcal{T}_h$ and $q \in P_K$. Using Lemma 60.1, we infer that

$$(u, \tilde{A}(q))_{L(K)} + \sum_{F \in \mathcal{F}_K} \epsilon_{K,F} (\Phi_F(u), q)_{L(F)} = (f, q)_{L(K)}, \tag{60.8}$$

where the flux function is defined by $\Phi_F(u) := \mathcal{N}_F u_{|F}$ for all $F \in \mathcal{F}_h$. Notice that the flux function is a notion attached to the mesh faces and not to the mesh cells. Then the local dG formulation with *centered fluxes* consists of seeking a discrete solution $u_h \in V_h$ such that

$$(u_h, \tilde{A}(q))_{L(K)} + \sum_{F \in \mathcal{F}_K} \epsilon_{K,F} (\widehat{\Phi}_F^{\rm cnt}(u_h), q)_{L(F)} = (f, q)_{L(K)}, \tag{60.9}$$

for all $K \in \mathcal{T}_h$ and $q \in P_K$, where the *centered numerical flux* is defined by

$$\widehat{\Phi}_F^{\rm cnt}(u_h) := \begin{cases} \mathcal{N}_F\{u_h\} & \text{if } F \in \mathcal{F}_h^\circ, \\ \frac{1}{2}(\mathcal{M}_F^{\rm BP} + \mathcal{N}_F) u_h & \text{if } F \in \mathcal{F}_h^\partial. \end{cases} \tag{60.10}$$

Notice that the centered flux is consistent with the exact flux in the sense that $\widehat{\Phi}_F^{\rm cnt}(u) = \Phi_F(u)$ for all $F \in \mathcal{F}_h^\circ$, since Corollary 60.2 implies that $\mathcal{N}_F\{u\} = \mathcal{N}_F u_{|F}$) and for all $F \in \mathcal{F}_h^\partial$ (since $(\mathcal{M}_F - \mathcal{N}_F) u_{|\partial D} = 0$ implies that $(\mathcal{M}_F^{\rm BP} - \mathcal{N}_F) u_{|\partial D} = 0$ owing to (57.33a).

Summing (60.9) over the cells in $\mathcal{T}_h$, we are lead to define the following sesquilinear form on $V_h \times V_h$:

$$a_h^{\rm cnt}(v_h, w_h) := (v_h, \tilde{A}_h(w_h))_L + \frac{1}{2}((\mathcal{M}^{\rm BP} + \mathcal{N}) v_h, w_h)_{L(\partial D)} + \overline{n_h(w_h, v_h)}. \tag{60.11}$$

Owing to (60.6), the discrete sesquilinear form $a_h^{\rm cnt}$ can also be rewritten as

$$a_h^{\rm cnt}(v_h, w_h) = (A_h(v_h), w_h)_L + \frac{1}{2}((\mathcal{M}^{\rm BP} - \mathcal{N}) v_h, w_h)_{L(\partial D)} - n_h(v_h, w_h). \tag{60.12}$$

The local problems (60.9) are then recast into the following global problem:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h^{\rm cnt}(u_h, w_h) = \ell_h(w_h) := (f, w_h)_L, \quad \forall w_h \in V_h. \end{cases} \tag{60.13}$$

60.2.2 Error analysis

We perform the error analysis using Lemma 27.8: we establish stability and consistency/boundedness, and we prove convergence by using the approximation properties of finite elements. Let us start with stability which takes the simple form of coercivity. Recall the seminorm $|v|_{\mathcal{M}^{\rm BP}} := (\mathcal{M}^{\rm BP} v, v)_{L(\partial D)}^{\frac{1}{2}}$.

Lemma 60.3 (Coercivity, well-posedness). (i) *The following holds true:*

$$\Re\big(a_h^{\mathrm{cnt}}(v_h,v_h)\big) \ge \mu_0\|v_h\|_L^2 + \frac{1}{2}|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 =: \|v_h\|_{V_h}^2, \tag{60.14}$$

for all $v_h \in V_h$. (ii) *The discrete problem* (60.13) *is well-posed.*

Proof. We only need to establish (60.14) since the well-posedness of (60.13) then follows from the Lax–Milgram lemma. We take the arithmetic mean of (60.11) and (60.12). Since $\overline{n_h(v_h,v_h)} = n_h(v_h,v_h)$ and since $(A_h(v_h),v_h)_L + (v_h,\tilde{A}_h(v_h))_L = (\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})v_h,v_h)_L$ is real, we infer that $\Re(a_h^{\mathrm{cnt}}(v_h,v_h)) = \frac{1}{2}((\mathcal{K}+\mathcal{K}^{\mathsf{H}}-\mathcal{X})v_h,v_h)_L + \frac{1}{2}|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2$. Then (60.14) follows from (56.1c). □

We assume that $\max(\|\mathcal{K}\|_{L^\infty(D;\mathbb{C}^{m\times m})},\|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}) \le c_{\mathcal{K},\mathcal{X}}\mu_0$ (see (57.10)), and for simplicity we hide the factor $c_{\mathcal{K},\mathcal{X}}$ in the generic constants used in the error analysis. As in the previous chapters, we set

$$\beta_K := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(K;\mathbb{C}^{m\times m})}, \qquad \beta := \max_{K\in\mathcal{T}_h}\beta_K. \tag{60.15}$$

We assume that the solution to (57.1) is in V_{S} with V_{S} defined in (60.7). We set $V_\sharp := V_{\mathrm{S}} + V_h$ and equip the space $V_\sharp$ with the following two norms:

$$\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac{1}{2}|v|_{\mathcal{M}^{\mathrm{BP}}}^2, \tag{60.16a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h}\mu_0^{-1}\beta_K^2(h_K^{-1}\|v\|_{L(\partial K)}^2 + \|\nabla v\|_{\boldsymbol{L}(K)}^2). \tag{60.16b}$$

Notice that (27.7) is satisfied with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ on $V_\sharp$).

Lemma 60.4 (Consistency/boundedness). *Let the consistency error be defined by*

$$\langle\delta_h(v_h),w_h\rangle_{V_h',V_h} := \ell_h(w_h) - a_h^{\mathrm{cnt}}(v_h,w_h), \qquad \forall v_h,w_h\in V_h.$$

There is $\omega_\sharp$, uniform w.r.t. $u\in V_{\mathrm{S}}$, such that for all $v_h,w_h\in V_h$ and all $h\in\mathcal{H}$,

$$|\langle\delta_h(v_h),w_h\rangle_{V_h',V_h}| \le \omega_\sharp\,\|u-v_h\|_{V_\sharp}\|w_h\|_{V_h}. \tag{60.17}$$

Proof. Since $A_h(u) = A(u) = f$ in L, $(\mathcal{M}^{\mathrm{BP}}-\mathcal{N})u = 0$ in $L(\partial D)$, and $\mathcal{N}_F[\![u]\!]_F = 0$ by Corollary 60.2 (so that $n_h(u,w_h)=0$), (60.12) implies that

$$\langle\delta_h(v_h),w_h\rangle_{V_h',V_h} = (A_h(\eta),w_h)_L - \frac{1}{2}((\mathcal{M}^{\mathrm{BP}}-\mathcal{N})\eta,w_h)_{L(\partial D)} + n_h(\eta,w_h),$$

with $\eta := u - v_h$. Let us bound the three terms composing the right-hand side, say $\mathfrak{T}_1,\mathfrak{T}_2,\mathfrak{T}_3$. Using the Cauchy–Schwarz inequality and the bound $\|A_h(\eta)\|_L \le c(\mu_0\|\eta\|_L + (\sum_{K\in\mathcal{T}_h}\beta_K^2\|\nabla\eta\|_{\boldsymbol{L}(K)}^2)^{\frac12})$, we infer that $|\mathfrak{T}_1| \le c\|\eta\|_{V_\sharp}\|w_h\|_{V_h}$. Using (57.33c), (57.33b), and a discrete trace inequality to bound $\|w_h\|_{L(F)}$ for all $F\in\mathcal{F}_h^\partial$, we also infer that $|\mathfrak{T}_2| \le c\|\eta\|_{V_\sharp}\|w_h\|_{V_h}$. The bound on $\mathfrak{T}_3$ is similar once the jumps and averages are bounded by a triangle inequality (i.e., $|[\![\eta]\!]_F| \le |\eta_{|K_l}| + |\eta_{|K_r}|$ with $F := \partial K_l\cap\partial K_r$, and so on) yielding $|\mathfrak{T}_3| \le c\sum_{K\in\mathcal{T}_h}\beta_K\|\eta\|_{L(\partial K)}\|w_h\|_{L(\partial K)}$. □

Theorem 60.5 (Error estimate). *Let u solve* (57.1) *and assume that $u\in V_{\mathrm{S}}$.* (i) *There is c s.t. for all $h\in\mathcal{H}$,*

$$\|u-u_h\|_{V_\flat} \le c\inf_{v_h\in V_h}\|u-v_h\|_{V_\sharp}. \tag{60.18}$$

(ii) *If* $u \in H^{1+r}(D;\mathbb{C}^m)$, $r \in [0,k]$, *then*

$$\|u-u_h\|_{V_\flat} \le c\left(\sum_{K\in\mathcal{T}_h} \phi_K h_K^{2r}|u|^2_{H^{1+r}(K;\mathbb{C}^m)}\right)^{\frac12}, \tag{60.19}$$

with $\phi_K := \max(\mu_0 h_K^2, \beta_K h_K, \mu_0^{-1}\beta_K^2)$.

Proof. Invoking Lemma 27.8 together with the above stability and consistency/boundedness results yields (60.18). Then (60.19) follows from the approximation properties of V_h (see (60.2)). □

The convergence result of Theorem 60.5 is suboptimal by one order in the L-norm and does not convey information on the convergence of derivatives (see the definition of the $V_\flat$-norm in (60.16a)). Moreover, no convergence result is achieved for $k=0$. We address these issues in §60.3.

60.2.3 Examples

Example 60.6 (Advection-reaction). Consider the PDE $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$ with the inflow boundary condition $u=0$ on ∂D^-; see §56.2.1. Assume that all the boundary faces of the mesh are subsets of either ∂D^- or $\partial D\backslash\partial D^-$. Then $\widehat{\Phi}_F^{\mathrm{cnt}}(u_h) := (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)\{u_h\}$ for all $F\in\mathcal{F}_h^\circ$ and $\widehat{\Phi}_F^{\mathrm{cnt}}(u_h) := \frac12(\boldsymbol{\beta}\cdot\boldsymbol{n}_F + |\boldsymbol{\beta}\cdot\boldsymbol{n}_F|)u_h$ for all $F\in\mathcal{F}_h^\partial$. The estimate of Theorem 60.5 with $r:=k$ gives $\mu_0^{\frac12}\|u-u_h\|_{L^2(D)} \le c\phi^{\frac12}h^k|u|_{H^{k+1}(D)}$ with $\mu_0 := \operatorname{ess\,inf}_D(\mu - \frac12\nabla\cdot\boldsymbol{\beta})$, $\phi := \max(\mu_0 h^2, \beta h, \mu_0^{-1}\beta^2)$, and $\beta := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$. □

Example 60.7 (Darcy). Consider the PDEs $\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \boldsymbol{0}$ and $\mu p + \nabla\cdot\boldsymbol{\sigma} = f$ with the boundary condition $p=0$; see §56.2.2. Recalling Example 57.12 and the scaling argument from §57.3.3, we introduce the scaling matrix defined in (57.25) with the two reference scales d_* and μ_* (e.g., $d_* := \lambda_\sharp$, $\mu_* := \mu_\flat$). The L-coercivity constant is $\mu_0 := \min(\frac{\mu_\flat}{\mu_*}, \frac{d_*}{\lambda_\sharp})$, and (60.15) gives $\beta_K = \ell_*$ with $\ell_* := (d_*/\mu_*)^{\frac12}$. Recalling the boundary penalty matrix $\mathcal{S}_F^\partial$ defined in (57.34), we have $\widehat{\Phi}_F^{\mathrm{cnt}}(\boldsymbol{\sigma}_h, p_h) := (\{p_h\}\boldsymbol{n}_F, \{\boldsymbol{\sigma}_h\}\cdot\boldsymbol{n}_F)$ for all $F\in\mathcal{F}_h^\circ$ and $\widehat{\Phi}_F^{\mathrm{cnt}}(\boldsymbol{\sigma}_h, p_h) := (0, \boldsymbol{\sigma}_h\cdot\boldsymbol{n}_F + \alpha_F p_h)$ for all $F\in\mathcal{F}_h^\partial$, where $\alpha_F := \alpha_*\beta_{K_l}\mu_*$ with a user-defined $\mathcal{O}(1)$ nondimensional parameter $\alpha_*>0$. Letting $\phi := \max(\mu_0 h^2, \ell_* h, \mu_0^{-1}\ell_*^2)$, the error estimate of Theorem 60.5 with $r:=k$ gives

$$\mu_0^{\frac12}\big(d_*^{-\frac12}\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} + \mu_*^{\frac12}\|p-p_h\|_{L^2(D)}\big) \le c\,\phi^{\frac12}h^k\big(d_*^{-\frac12}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+1}(D)} + \mu_*^{\frac12}|p|_{H^{k+1}(D)}\big).$$ □

Example 60.8 (Maxwell). Consider the PDEs $\sigma\boldsymbol{E} - \nabla\times\boldsymbol{H} = \boldsymbol{f}$ and $\mathrm{i}\omega\mu\boldsymbol{H} + \nabla\times\boldsymbol{E} = \boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$. Recalling Example 57.13 and the scaling argument from §57.3.3, we introduce the reference scales σ_* and $\tilde{\mu}_*$ (e.g., $\sigma_* := \sigma_\flat$, $\widetilde{\mu}_* := \omega\mu_\flat$). The L-coercivity constant is $\mu_0 := \frac{1}{\sqrt{2}}\min(\frac{\sigma_\flat}{\sigma_*}, \frac{\omega\mu_\flat}{\widetilde{\mu}_*})$, and (60.15) gives $\beta_K = \ell_*$ where $\ell_* := (\sigma_*\tilde{\mu}_*)^{-\frac12}$. Recalling the boundary penalty matrix $\mathcal{S}_F^\partial$ defined in (57.35), we have $\widehat{\Phi}_F^{\mathrm{cnt}}(\boldsymbol{E}_h, \boldsymbol{H}_h) := (\{\boldsymbol{E}_h\}\times\boldsymbol{n}_F, \{\boldsymbol{H}_h\}\times\boldsymbol{n}_F)$ for all $F\in\mathcal{F}_h^\circ$ and $\widehat{\Phi}_F^{\mathrm{cnt}}(\boldsymbol{E}_h, \boldsymbol{H}_h) := (\boldsymbol{E}_h\times\boldsymbol{n}_F + \alpha_F\boldsymbol{n}_F\times(\boldsymbol{H}_h\times\boldsymbol{n}_F), \boldsymbol{0})$ for all $F\in\mathcal{F}_h^\partial$, where $\alpha_F := \alpha_*\beta_{K_l}\widetilde{\mu}_*$ with a user-defined $\mathcal{O}(1)$ nondimensional parameter $\alpha_*>0$. Letting $\phi := \max(\mu_0 h^2, \ell_* h, \mu_0^{-1}\ell_*^2)$, the error estimate of Theorem 60.5 with $r:=k$ gives

$$\mu_0^{\frac12}\big(\sigma_*^{\frac12}\|\boldsymbol{E}-\boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)} + \tilde{\mu}_*^{\frac12}\|\boldsymbol{H}-\boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)}\big) \le c\,\phi^{\frac12}h^k\big(\sigma_*^{\frac12}|\boldsymbol{E}|_{\boldsymbol{H}^{k+1}(D)} + \tilde{\mu}_*^{\frac12}|\boldsymbol{H}|_{\boldsymbol{H}^{k+1}(D)}\big).$$ □

60.3 Tightened stability by jump penalty

In this section, we improve on the shortcomings of the centered numerical flux by tightening the stability properties of the discrete sesquilinear form.

60.3.1 Local and global formulation

The key idea is to add to the centered-flux-based sesquilinear form $a_h^{\rm cnt}$ a stabilization term penalizing the interface jumps. We then set

$$a_h^{\rm stb}(v_h,w_h) := a_h^{\rm cnt}(v_h,w_h) + \sum_{F\in\mathcal{F}_h^\circ} (\mathcal{S}_F^\circ[\![v_h]\!],[\![w_h]\!])_{L(F)}, \tag{60.20}$$

where the interface penalty field $\mathcal{S}_F^\circ$ is Hermitian and positive semidefinite for all $F\in\mathcal{F}_h^\circ$. Notice that the boundary conditions are accounted for by $a_h^{\rm cnt}$ which incorporates the contribution of the boundary penalty method. We define the seminorm $|v|_{\mathcal{S}_F^\circ} := (\mathcal{S}_F^\circ v,v)_{L(F)}^{\frac12}$. The above assumptions on $\mathcal{S}_F^\circ$ imply that $|(\mathcal{S}_F^\circ v,w)_{L(F)}|\le |v|_{\mathcal{S}_F^\circ}|w|_{\mathcal{S}_F^\circ}$ for all $v,w\in L(F)$. Moreover, we assume that there is c s.t. for all $h\in\mathcal{H}$,

$$\ker(\mathcal{N}_F)\subset\ker(\mathcal{S}_F^\circ), \tag{60.21a}$$

$$|v|_{\mathcal{S}_F^\circ}\le c\,\beta_F^{\frac12}\|v\|_{L(F)}, \tag{60.21b}$$

$$|(\mathcal{N}_F v,w)_{L(F)}|\le c\,|v|_{\mathcal{S}_F^\circ}\beta_F^{\frac12}\|w\|_{L(F)}, \tag{60.21c}$$

with $\beta_F := \|\mathcal{N}_F\|_{L^\infty(F;\mathbb{C}^{m\times m})}$. Notice that $\beta_F\le c\min(\beta_{K_l},\beta_{K_r})$ with $F:=\partial K_l\cap\partial K_r$. The discrete problem is formulated as follows:

$$\left\{\begin{array}{l}\text{Find } u_h\in V_h \text{ such that}\\ a_h^{\rm stb}(u_h,w_h)=\ell_h(w_h):=(f,w_h)_L,\quad \forall w_h\in V_h.\end{array}\right. \tag{60.22}$$

Let us define the stabilized *numerical flux* (compare with (60.10))

$$\widehat{\Phi}_F^{\rm stb}(u_h) = \begin{cases}\mathcal{N}_F\{u_h\}+\mathcal{S}_F^\circ[\![u_h]\!] & \text{if } F\in\mathcal{F}_h^\circ,\\ \frac12(\mathcal{M}_F^{\rm BP}+\mathcal{N}_F)u_h & \text{if } F\in\mathcal{F}_h^\partial.\end{cases} \tag{60.23}$$

Then u_h solves (60.22) iff u_h is s.t. for all $K\in\mathcal{T}_h$ and all $q\in P_K$,

$$(u_h,\tilde{A}(q))_{L(K)} + \sum_{F\in\mathcal{F}_K}\epsilon_{K,F}(\widehat{\Phi}_F^{\rm stb}(u_h),q)_{L(F)} = (f,q)_{L(K)}. \tag{60.24}$$

60.3.2 Error analysis

The crucial improvement with respect to the formulation using centered fluxes is that we can now establish inf-sup stability with the following stronger norm:

$$\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac12|v_h|_{\mathcal{M}^{\rm BP}}^2 + |[\![v_h]\!]|_{\mathcal{S}^\circ}^2 + \|\tau^{\frac12}A_{1h}(v_h)\|_L^2, \tag{60.25}$$

with the jump seminorm $|[\![v_h]\!]|_{\mathcal{S}^\circ} := (\sum_{F\in\mathcal{F}_h^\circ}|[\![v_h]\!]_F|_{\mathcal{S}_F^\circ}^2)^{\frac12}$, and the piecewise constant function τ such that $\tau_{|K}:=\tau_K$ for all $K\in\mathcal{T}_h$ with the local weights τ_K defined in (58.2):

$$\tau_K := \big(\max(\beta_K h_K^{-1},\mu_0)\big)^{-1} = \min(\beta_K^{-1}h_K,\mu_0^{-1}). \tag{60.26}$$

We assume that $\mathcal{A}_{|K}^k\in C^{0,\frac12}(K;\mathbb{C}^{m\times m})$ for all $k\in\{1{:}d\}$ and all $K\in\mathcal{T}_h$. Letting $\underline{\mathcal{A}}_K^k := |K|^{-1}\int_K\mathcal{A}^k\,{\rm d}x$, there is $c_{\mathcal{A}}$ s.t.

$$\|\mathcal{A}^k-\underline{\mathcal{A}}_K^k\|_{L^\infty(K;\mathbb{C}^{m\times m})}\le c_{\mathcal{A}}\,(\mu_0\beta_K h_K)^{\frac12}, \tag{60.27}$$

and we hide the factor $c_{\mathcal{A}}$ in the generic constants c used in the error analysis.

Lemma 60.9 (Stability, well-posedness). (i) *There is $\alpha > 0$ such that for all $h \in \mathcal{H}$,*

$$\inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{|a_h^{\mathrm{stb}}(v_h, w_h)|}{\|v_h\|_{V_h} \|w_h\|_{V_h}} \geq \alpha > 0. \tag{60.28}$$

(ii) *The discrete problem* (60.22) *is well-posed.*

Proof. We only need to establish (60.28) since the well-posedness of (60.22) directly follows from (60.28). Let $v_h \in V_h$ and set $r_h := \sup_{w_h \in V_h} \frac{|a_h^{\mathrm{stb}}(v_h, w_h)|}{\|w_h\|_{V_h}}$. Our goal is to prove that there is $\alpha > 0$ s.t. $\alpha \|v_h\|_{V_h} \leq r_h$ for all $h \in \mathcal{H}$.
(1) Owing to the coercivity of a_h^{cnt} (see (60.14)) and by definition of the jump seminorm, we have

$$\mu_0 \|v_h\|_L^2 + \frac{1}{2} |v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v_h]\!]|_{\mathcal{S}^\circ}^2 \leq \frac{|a_h^{\mathrm{stb}}(v_h, v_h)|}{\|v_h\|_{V_h}} \|v_h\|_{V_h} \leq r_h \|v_h\|_{V_h}.$$

(2) Let $\underline{A}_{1h}(v_h)$ be such that $\underline{A}_{1h}(v_h)_{|K} := \sum_{k \in \{1:d\}} \underline{A}_K^k \partial_k v_{h|K}$ for all $K \in \mathcal{T}_h$. Set $w_h := \tau \underline{A}_{1h}(v_h)$ and observe that $w_h \in V_h$. The triangle inequality and the definition of the $\|\cdot\|_{V_h}$-norm imply that

$$\|\tau^{-\frac{1}{2}} w_h\|_L \leq \|\tau^{\frac{1}{2}} (\underline{A}_{1h} - A_{1h})(v_h)\|_L + \|v_h\|_{V_h}.$$

Using (60.27), an inverse inequality, and the definition of τ_K, we infer that $\|\tau^{\frac{1}{2}} (\underline{A}_{1h} - A_{1h})(v_h)\|_L \leq c \mu_0^{\frac{1}{2}} \|v_h\|_L$. Therefore, we have $\|\tau^{-\frac{1}{2}} w_h\|_L \leq c \|v_h\|_{V_h}$. Furthermore, proceeding as in the proof of Lemma 58.2, one proves that $\mu_0^{\frac{1}{2}} \|y_h\|_L + |y_h|_{\mathcal{M}^{\mathrm{BP}}} + \|\tau^{\frac{1}{2}} A_{1h}(y_h)\|_L \leq c \|\tau^{-\frac{1}{2}} y_h\|_L$ for all $y_h \in V_h$. Owing to (60.21b), (60.27), and using a discrete trace inequality, we also infer that $|[\![y_h]\!]|_{\mathcal{S}^\circ} \leq c \|\tau^{-\frac{1}{2}} y_h\|_L$. Applying these bounds to $y_h := w_h$ yields

$$\|w_h\|_{V_h} + \|\tau^{-\frac{1}{2}} w_h\|_L \leq c \, \|v_h\|_{V_h}. \tag{60.29}$$

(3) Using the expression (60.12) for a_h^{cnt}, we observe that

$$\begin{aligned} \|\tau^{\frac{1}{2}} A_{1h}(v_h)\|_L^2 = \|\tau^{\frac{1}{2}} A_{1h}(v_h)\|_L^2 = a_h^{\mathrm{stb}}(v_h, w_h) - (\mathcal{K} v_h, w_h)_L + (\tau A_{1h} v_h, (A_{1h} - \underline{A}_{1h})(v_h))_L \\ - \frac{1}{2} ((\mathcal{M}^{\mathrm{BP}} - \mathcal{N}) v_h, w_h)_{L(\partial D)} + n_h(v_h, w_h) - \sum_{F \in \mathcal{F}_h^\circ} (\mathcal{S}_F^\circ [\![v_h]\!], [\![w_h]\!])_{L(F)}. \end{aligned}$$

Let $\mathfrak{T}_1, \dots, \mathfrak{T}_6$ be the terms on the right-hand side. Owing to (60.29), we have $|\mathfrak{T}_1| \leq r_h \|w_h\|_{V_h} \leq c \, r_h \|v_h\|_{V_h}$. By proceeding as in the proof of Lemma 58.2 to bound $\mathfrak{T}_2 + \mathfrak{T}_4$ and by using the Cauchy–Schwarz inequality $|(\mathcal{S}_F^\circ v, w)_{L(F)}| \leq |v|_{\mathcal{S}_F^\circ} |w|_{\mathcal{S}_F^\circ}$, we obtain

$$|\mathfrak{T}_2 + \mathfrak{T}_4 + \mathfrak{T}_6| \leq c \, r_h^{\frac{1}{2}} \|v_h\|_{V_h}^{\frac{1}{2}} \|w_h\|_{V_h} \leq c' \, r_h^{\frac{1}{2}} \|v_h\|_{V_h}^{\frac{3}{2}},$$

where we used again (60.29). Employing (60.21c) gives $|\mathfrak{T}_5| \leq c \sum_{F \in \mathcal{F}_h^\circ} |[\![v_h]\!]|_{\mathcal{S}_F^\circ} \beta_F^{\frac{1}{2}} \|\{w_h\}\|_{L(F)}$. The triangle inequality and the bound $\beta_F \leq c \min(\beta_{K_l}, \beta_{K_r})$ imply that

$$\beta_F^{\frac{1}{2}} \|\{w_h\}\|_{L^2(F)} \leq c \sum_{K \in \mathcal{T}_F} \beta_K^{\frac{1}{2}} \|w_{h|K}\|_{L^2(F)}.$$

Using a discrete trace inequality and $\beta_K h_K^{-1} \leq \tau_K^{-1}$, we have

$$\beta_F^{\frac{1}{2}} \|\{w_h\}\|_{L^2(F)} \leq c \sum_{K \in \mathcal{T}_F} \tau_K^{-\frac{1}{2}} \|w_h\|_{L^2(K)}.$$

The bound (60.29) on w_h and the Cauchy–Schwarz inequality yield $|\mathfrak{T}_5| \le c r_h^{\frac12} \|v_h\|_{V_h}^{\frac32}$. Employing Young's inequality and the above bound on $\|\tau^{\frac12}(\underline{A}_{1h} - A_{1h})(v_h)\|_L$ gives

$$|\mathfrak{T}_3| \le \frac12 \|\tau^{\frac12} A_{1h}(v_h)\|_L^2 + c\,\mu_0 \|v_h\|_L^2 \le \frac12 \|\tau^{\frac12} A_{1h}(v_h)\|_L^2 + c\, r_h \|v_h\|_{V_h}.$$

(4) Collecting the above bounds yields $\|v_h\|_{V_h}^2 \le c r_h^{\frac12} \|v_h\|_{V_h}^{\frac32} + r_h \|v_h\|_{V_h}$, and we conclude the proof of (60.28). □

As we did when we analyzed the method with centered fluxes, we assume that the solution to the model problem (57.1) is such that

$$u \in V_{\mathrm{s}} := H^s(D;\mathbb{C}^m) \cap V, \qquad s > \frac12. \tag{60.30}$$

We set $V_\sharp := V_{\mathrm{s}} + V_h$, and we equip the space $V_\sharp$ with the following two norms:

$$\|v\|_{V_\flat}^2 := \mu_0 \|v\|_L^2 + \frac12 |v|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v]\!]|_{\mathcal{S}^\circ}^2 + \|\tau^{\frac12} A_{1h}(v)\|_L^2, \tag{60.31a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h} \Big(\tau_K^{-1} \|v\|_{L(K)}^2 + \beta_K \|v\|_{L(\partial K)}^2 \Big), \tag{60.31b}$$

so that (27.7) is satisfied with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \le \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \le \|v\|_{V_\sharp}$ on $V_\sharp$).

Lemma 60.10 (Consistency/boundedness). *Define the consistency error as*

$$\langle \delta_h(v_h), w_h \rangle_{V_h',V_h} := \ell_h(w_h) - a_h^{\mathrm{stb}}(v_h, w_h), \qquad \forall v_h, w_h \in V_h.$$

There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{s}}$, s.t. for all $v_h, w_h \in V_h$, all $h \in \mathcal{H}$,

$$|\langle \delta_h(v_h), w_h \rangle_{V_h',V_h}| \le \omega_\sharp \, \|u - v_h\|_{V_\sharp} \|w_h\|_{V_h}. \tag{60.32}$$

Proof. Using the same arguments as in the proof of Lemma 60.4, but using now the expression (60.11) for a_h^{cnt}, we obtain

$$\begin{aligned}\langle \delta_h(v_h), w_h \rangle_{V_h',V_h} &= (\eta, \tilde{A}_h(w_h))_L + \frac12 ((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta, w_h)_{L(\partial D)} + \overline{n_h(w_h,\eta)} \\ &\quad + \sum_{F\in\mathcal{F}_h^\circ} (\mathcal{S}_F^\circ [\![\eta]\!], [\![w_h]\!])_{L(F)} =: \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3 + \mathfrak{T}_4,\end{aligned}$$

with $\eta := u - v_h$. The terms $\mathfrak{T}_1$ and $\mathfrak{T}_2$ can be bounded as in the proof of Lemma 57.21. Proceeding as in the proof of Lemma 60.9 (bound on $\mathfrak{T}_5$) yields $|\mathfrak{T}_3| \le c(\sum_{K\in\mathcal{T}_h} \beta_K \|v\|_{L(\partial K)}^2)^{\frac12} |[\![w_h]\!]|_{\mathcal{S}^\circ} \le c\|v\|_{V_\sharp} \|w_h\|_{V_h}$. Finally, $|\mathfrak{T}_4| \le |v|_{\mathcal{S}^\circ} |w_h|_{\mathcal{S}^\circ} \le \|v\|_{V_\sharp} \|w_h\|_{V_h}$. □

Theorem 60.11 (Error estimate). *Let u solve (57.1) and assume $u \in V_{\mathrm{s}}$.* (i) *There is c s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\flat} \le c \inf_{v_h\in V_h} \|u - v_h\|_{V_\sharp}. \tag{60.33}$$

(ii) *If $u \in H^{1+r}(D;\mathbb{C}^m)$, $r \in [0,k]$, then*

$$\|u - u_h\|_{V_\flat} \le c \left(\sum_{K\in\mathcal{T}_h} (\tau_K^{-1} h_K) h_K^{2r+1} |u|_{H^{1+r}(K;\mathbb{C}^m)}^2 \right)^{\frac12}, \tag{60.34}$$

i.e., $\|u - u_h\|_{V_\flat} \le c\phi^{\frac12} h^{r+\frac12} |u|_{H^{1+r}(D;\mathbb{C}^m)}$ with $\phi := \max(\max_{K\in\mathcal{T}_h} \beta_K, \mu_0 h)$.

Proof. Similar to that of Theorem 60.5. □

Remark 60.12 (Literature). The analysis of dG methods for Friedrichs' systems started in the 1970s with Lesaint [214], Lesaint and Raviart [215] and was later refined by Johnson and Pitkäranta [200]. A systematic treatment was given in [118, 119, 120]; see also Jensen [197]. The devising of dG methods with tightened stability by means of a jump penalty is found in Brezzi et al. [54] for advection-reaction. □

60.3.3 Examples

Let ∇_h denote the broken gradient operator (see Definition 36.3).

Example 60.13 (Advection-reaction). Recalling Example 60.6, we consider the PDE $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$ with the inflow boundary condition $u = 0$ on ∂D^-. The jump penalty coefficient can be set to $\mathcal{S}_F^\circ := \alpha_*|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|$ for all $F \in \mathcal{F}_h^\circ$, where $\alpha_* > 0$ is a user-defined $\mathcal{O}(1)$ nondimensional parameter. In other words, the jump of u_h is penalized across all the mesh interfaces where $|\boldsymbol{\beta}\cdot\boldsymbol{n}_F| > 0$. The numerical flux obtained by setting $\alpha_* := \frac{1}{2}$ is usually called *upwind flux* in the literature; see Exercise 60.1. We refer the reader to Burman and Stamm [70], Burman et al. [75] for further insight into the choice of the penalty parameter. Letting $\tau_{|K} := \min(\beta_K^{-1}h_K, \mu_0^{-1})$ with $\beta_K := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$ for all $K \in \mathcal{T}_h$, $\mu_0 := \operatorname{ess\,inf}_D(\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta})$, and $\phi := \max(\beta, \mu_0 h)$ with $\beta := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$, the error estimate from Theorem 60.11 (with $r := k$) gives

$$\mu_0^{\frac{1}{2}}\|u - u_h\|_{L^2(D)} + \|\tau^{\frac{1}{2}}\boldsymbol{\beta}\cdot\nabla_h(u - u_h)\|_{L^2(D)} \le c\,\phi^{\frac{1}{2}}h^{k+\frac{1}{2}}|u|_{H^{k+1}(D)}.$$ □

Example 60.14 (Darcy). We consider the PDEs $\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \boldsymbol{0}$ and $\mu p + \nabla\cdot\boldsymbol{\sigma} = f$ with the boundary condition $p = 0$. Recalling Example 60.7 and the reference scales d_* and μ_*, the L-coercivity constant is $\mu_0 := \min(\frac{\mu_\flat}{\mu_*}, \frac{d_*}{\lambda_\sharp})$, and (60.15) gives $\beta_K = \beta_F := \ell_*$ with $\ell_* := (d_*/\mu_*)^{\frac{1}{2}}$. The following jump and boundary penalty fields satisfy (60.21) and (57.33):

$$\mathcal{S}_F^\circ := \left[\begin{array}{c:c} \alpha_{1,F}\boldsymbol{n}_F\otimes\boldsymbol{n}_F & \mathbb{O}_{d\times 1} \\ \hdashline \mathbb{O}_{1\times d} & \alpha_{2,F} \end{array}\right], \qquad \mathcal{S}_F^\partial := \left[\begin{array}{c:c} \mathbb{O}_{d\times d} & \mathbb{O}_{d\times 1} \\ \hdashline \mathbb{O}_{1\times d} & \alpha_{2,F} \end{array}\right],$$

where $\alpha_{1,F} := \alpha_{1*}\beta_F d_*^{-1}$, $\alpha_{2,F} := \alpha_{2*}\beta_F\mu_*$, with user-defined $\mathcal{O}(1)$ nondimensional parameters $\alpha_{1*}, \alpha_{2*} > 0$. In other words, the jumps across the mesh interfaces of the normal component of $\boldsymbol{\sigma}_h$ and of p_h are penalized. Letting $\phi := \max(\ell_*, \mu_0 h)$ and recalling (60.26), we set $\tau_{|K} := \min(\ell_*^{-1}h_K, \mu_0^{-1})$ for all $K \in \mathcal{T}_h$. The error estimate from Theorem 60.11 (with $r := k$) gives

$$\begin{aligned}&\mu_0^{\frac{1}{2}}\big(d_*^{-\frac{1}{2}}\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} + \mu_*^{\frac{1}{2}}\|p - p_h\|_{L^2(D)}\big)\\ &\quad + \ell_* d_*^{-\frac{1}{2}}\|\tau^{\frac{1}{2}}\nabla_h\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)} + \ell_*\mu_*^{\frac{1}{2}}\|\tau^{\frac{1}{2}}\nabla_h(p - p_h)\|_{\boldsymbol{L}^2(D)}\\ &\quad\le c\,\phi^{\frac{1}{2}}h^{k+\frac{1}{2}}\big(d_*^{-\frac{1}{2}}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+1}(D)} + \mu_*^{\frac{1}{2}}|p|_{H^{k+1}(D)}\big).\end{aligned}$$ □

Example 60.15 (Maxwell). We consider the PDEs $\sigma\boldsymbol{E} - \nabla\times\boldsymbol{H} = \boldsymbol{f}$ and $\mathrm{i}\omega\mu\boldsymbol{H} + \nabla\times\boldsymbol{E} = \boldsymbol{0}$ with the boundary condition $\boldsymbol{H}\times\boldsymbol{n} = \boldsymbol{0}$. Recalling Example 60.8 and the reference scales σ_* and $\tilde{\mu}_*$, the L-coercivity constant is $\mu_0 := \frac{1}{\sqrt{2}}\min(\frac{\sigma_\flat}{\sigma_*}, \frac{\omega\mu_\flat}{\tilde{\mu}_*})$, and (60.15) gives $\beta_K = \beta_F := \ell_*$ with $\ell_* := (\sigma_*\tilde{\mu}_*)^{-\frac{1}{2}}$. The following jump and boundary penalty fields satisfy (60.21) and (57.33):

$$\mathcal{S}_F^\circ := \left[\begin{array}{c:c} \alpha_{1,F}\mathbb{T}_F^\intercal\mathbb{T}_F & \mathbb{O}_{3\times 3} \\ \hdashline \mathbb{O}_{3\times 3} & \alpha_{2,F}\mathbb{T}_F^\intercal\mathbb{T}_F \end{array}\right], \qquad \mathcal{S}_F^\partial := \left[\begin{array}{c:c} \mathbb{O}_{3\times 3} & \mathbb{O}_{3\times 3} \\ \hdashline \mathbb{O}_{3\times 3} & \alpha_{2,F}\mathbb{T}_F^\intercal\mathbb{T}_F \end{array}\right],$$

where $\alpha_{1,F} := \alpha_{1*}\beta_F\sigma_*$, $\alpha_{2,F} := \alpha_{2*}\beta_F\widetilde{\mu}_*$ with user-defined $\mathcal{O}(1)$ nondimensional parameters $\alpha_{1*}, \alpha_{2*} > 0$. In other words, the jumps across the mesh interfaces of the tangential components of $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$ are penalized. The matrix $\mathbb{T}_F$ is s.t. $\mathbb{T}_F\boldsymbol{\xi} := \boldsymbol{\xi}\times\boldsymbol{n}_F$ for all $\boldsymbol{\xi} \in \mathbb{C}^3$ (see §56.2.3). Letting $\phi := \max(\ell_*, \mu_0 h)$ and recalling (60.26), we set $\tau_{|K} := \min(\ell_*^{-1}h_K, \mu_0^{-1})$ for all $K \in \mathcal{T}_h$. The error estimate from Theorem 60.11 (with $r := k$) gives

$$\begin{aligned}&\mu_0^{\frac12}\big(\sigma_*^{\frac12}\|\boldsymbol{E}-\boldsymbol{E}_h\|_{\boldsymbol{L}^2(D)} + \widetilde{\mu}_*^{\frac12}\|\boldsymbol{H}-\boldsymbol{H}_h\|_{\boldsymbol{L}^2(D)}\big)\\ &\quad+\ell_*\sigma_*^{\frac12}\|\tau^{\frac12}\nabla_h\times(\boldsymbol{E}-\boldsymbol{E}_h)\|_{\boldsymbol{L}^2(D)} + \ell_*\widetilde{\mu}_*^{\frac12}\|\tau^{\frac12}\nabla_h\times(\boldsymbol{H}-\boldsymbol{H}_h)\|_{\boldsymbol{L}^2(D)}\\ &\quad\le c\,\phi^{\frac12}h^{k+\frac12}\big(\sigma_*^{\frac12}|\boldsymbol{E}|_{\boldsymbol{H}^{k+1}(D)} + \widetilde{\mu}_*^{\frac12}|\boldsymbol{H}|_{\boldsymbol{H}^{k+1}(D)}\big).\end{aligned}$$

We refer the reader to Houston et al. [187, 188] for further results on the dG approximation of the time-harmonic Maxwell's equations. □

Exercises

Exercise 60.1 (Upwind flux). Consider the advection equation $\mu u + \boldsymbol{\beta}\cdot\nabla u = f$. Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Let $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) := \boldsymbol{\beta}\cdot\boldsymbol{n}_F\{u_h\} + \frac12|\boldsymbol{\beta}\cdot\boldsymbol{n}_F|[\![u_h]\!]$. Show that $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) = (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)u_{h|K_l}$ if $\boldsymbol{\beta}\cdot\boldsymbol{n}_F \ge 0$ and $\widehat{\Phi}_F^{\mathrm{stb}}(u_h) = (\boldsymbol{\beta}\cdot\boldsymbol{n}_F)u_{h|K_r}$ otherwise.

Exercise 60.2 ($\mathcal{S}_F^\circ$). Verify that the jump penalty operators from §60.3.3 verify (60.21).

Exercise 60.3 (Absolute value). (i) Show that a suitable choice for the jump penalty operator is $\mathcal{S}_F^\circ = |\mathcal{N}_F|$ where $|\mathcal{N}_F|$ is the unique Hermitian positive semidefinite matrix such that $|\mathcal{N}_F|^2 = \mathcal{N}_F^{\mathsf{H}}\mathcal{N}_F = \mathcal{N}_F^2$. (*Hint*: $\big|w^{\mathsf{H}}\mathcal{N}_F v\big| \le \big|w^{\mathsf{H}}|\mathcal{N}_F|v\big|$.) (ii) Verify that

$$\left|\left[\begin{array}{c|c}\mathbb{O}_{d\times d} & \boldsymbol{n}_F\\ \hline \boldsymbol{n}_F^{\mathsf{T}} & 0\end{array}\right]\right| = \left[\begin{array}{c|c}\boldsymbol{n}_F\otimes\boldsymbol{n}_F & \boldsymbol{0}\\ \hline \boldsymbol{0}^{\mathsf{T}} & 1\end{array}\right],\quad \left|\left[\begin{array}{c|c}\alpha\mathbb{T}^{\mathsf{T}}\mathbb{T} & \mathbb{O}_{3\times3}\\ \hline \mathbb{O}_{3\times3} & \beta\mathbb{T}^{\mathsf{T}}\mathbb{T}\end{array}\right]\right| = \left[\begin{array}{c|c}|\alpha|\mathbb{T}^{\mathsf{T}}\mathbb{T} & \mathbb{O}_{3\times3}\\ \hline \mathbb{O}_{3\times3} & |\beta|\mathbb{T}^{\mathsf{T}}\mathbb{T}\end{array}\right].$$

Exercise 60.4 (Matrix $\mathbb{T}$). (i) Show that $\mathbb{T}^{\mathsf{T}} = -\mathbb{T}$. (ii) Show that $(\mathbb{T}^{\mathsf{T}}\mathbb{T})^2 = \mathbb{T}^{\mathsf{T}}\mathbb{T}$.

Exercise 60.5 (Orthogonal subscales). (i) Prove that a_h^{stb} is coercive on V_h equipped with the norm $\|v_h\|_{V_h}^2 := \mu_0\|v_h\|_L^2 + \frac12|v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v_h]\!]|_{\mathcal{S}^\circ}^2$. (ii) Assume that the fields $\mathcal{A}^k$ are Lipschitz (with Lipschitz constant $L_{\mathcal{A}} \le c\mu_0$). Assume that $u \in V_{\mathrm{s}} := H^s(D;\mathbb{C}^m)\cap V$, $s > \frac12$. Prove that there is c such that

$$|\langle\delta_h(\mathcal{I}_h^{\mathrm{b}}(u)), w_h\rangle_{V_h',V_h}| \le c\,\|u-\mathcal{I}_h^{\mathrm{b}}(u)\|_{V_\sharp}\|w_h\|_{V_h},$$

for all $(v, w_h) \in V_\sharp\times V_h$ and all $h \in \mathcal{H}$, where $\mathcal{I}_h^{\mathrm{b}}$ denotes the L-orthogonal projection onto V_h, $\|v\|_{V_\flat}^2 := \mu_0\|v\|_L^2 + \frac12|v|_{\mathcal{M}^{\mathrm{BP}}}^2 + |[\![v]\!]|_{\mathcal{S}^\circ}^2$, and $\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h}\beta_K\|v\|_{L(\partial K)}^2$. (*Hint*: adapt the proof of Lemma 60.10.) (iii) Prove that $\|u-u_h\|_{V_\flat} \le c\phi^{\frac12}h^{k+\frac12}|u|_{H^{k+1}(D;\mathbb{C}^m)}$ using only Steps (i) and (ii). (*Hint*: adapt the proof of Lemma 27.8.)

Chapter 61

Advection-diffusion

In this chapter, we want to solve a model problem where the PDE comprises a first-order differential operator modeling advection processes and a second-order term modeling diffusion processes. Advection-diffusion problems are encountered in many applications, e.g., heat transfer or pollutant transport by fluids, and constitute the first step toward the approximation of the Navier–Stokes equations. The difficulty in approximating an advection-diffusion equation can be quantified by the Péclet number which is equal to the meshsize times the advection velocity divided by the diffusion coefficient. When the Péclet number is small, i.e., when the mesh is fine enough, the problem can be approximated by the standard Galerkin method using H^1-conforming finite elements as done in Chapter 32 for the pure diffusion problem. But when the Péclet number is large, the standard Galerkin approximation is plagued by spurious oscillations. These oscillations disappear if very fine meshes are used, but a more effective approach using coarser meshes is to resort to stabilization. In this chapter, we focus on the Galerkin/least-squares (GaLS) stabilization, but any stabilized H^1-conforming method or the dG method can also be used. More generally, the advection-diffusion problem is a prototype for studying *singularly perturbed* elliptic PDEs.

61.1 Model problem

Let D be a Lipschitz domain in $\mathbb{R}^d$ and let $f \in L^2(D)$. The model problem we want to approximate is as follows:

$$T_\epsilon(u) := -\nabla\cdot(\mathbb{d}_\epsilon \nabla u) + A(u) = f \quad \text{in } D, \tag{61.1a}$$

$$u = 0 \quad \text{on } \partial D, \tag{61.1b}$$

with the diffusion tensor $\mathbb{d}_\epsilon \in \mathbb{L}^\infty(D) := L^\infty(D;\mathbb{R}^{d\times d})$ taking symmetric positive definite values. We assume that the smallest eigenvalue of $\mathbb{d}_\epsilon$ is uniformly bounded away from zero by a real number $\epsilon > 0$, and the first-order operator A is defined by $A(u) := \boldsymbol{\beta}\cdot\nabla u + \mu u$ with $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D) := W^{1,\infty}(D;\mathbb{R}^d)$, $\mu \in L^\infty(D)$, and $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \ge \mu_0 > 0$ a.e. in D; see §56.2.1. We focus on homogeneous Dirichlet boundary conditions, but any of the boundary conditions considered in Chapter 31 can be considered. Setting $V := H^1_0(D)$, a weak formulation of (61.1) is the following:

$$\begin{cases} \text{Find } u \in V := H^1_0(D) \text{ such that} \\ a_\epsilon(u,w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{61.2}$$

where

$$a_\epsilon(v,w) := (\mathbb{d}_\epsilon \nabla v, \nabla w)_{\boldsymbol{L}^2(D)} + (A(v),w)_{L^2(D)}, \tag{61.3a}$$

$$\ell(w) := (f,w)_{L^2(D)}. \tag{61.3b}$$

Lemma 61.1 (Well-posedness, a priori estimates). (i) *The bilinear form a_ϵ is coercive on V:*

$$a_\epsilon(v,v) \geq \epsilon\,|v|^2_{H^1(D)} + \mu_0 \|v\|^2_{L^2(D)}, \qquad \forall v \in V. \tag{61.4}$$

(ii) *The problem* (61.2) *is well-posed.* (iii) *The solution satisfies the a priori estimates*

$$\|u\|_{L^2(D)} \leq \mu_0^{-1} \|f\|_{L^2(D)}, \qquad |u|_{H^1(D)} \leq (4\mu_0\epsilon)^{-\frac{1}{2}} \|f\|_{L^2(D)}. \tag{61.5}$$

Proof. The coercivity property (61.4) follows from the assumptions on $\mathbb{d}_\epsilon$ and A. The well-posedness of (61.2) results from the Lax–Milgram lemma. Let us now establish the a priori estimates in (61.5). We observe that

$$\epsilon\,|u|^2_{H^1(D)} + \mu_0 \|u\|^2_{L^2(D)} \leq \|f\|_{L^2(D)} \|u\|_{L^2(D)}. \tag{61.6}$$

Thus, $\mu_0 \|u\|^2_{L^2(D)} \leq \|f\|_{L^2(D)} \|u\|_{L^2(D)}$, and this yields the bound on $\|u\|_{L^2(D)}$. Moreover, the inequality $\|f\|_{L^2(D)} \|u\|_{L^2(D)} \leq \frac{1}{4}\mu_0^{-1} \|f\|^2_{L^2(D)} + \mu_0 \|u\|^2_{L^2(D)}$ combined with (61.6) implies that $\epsilon |u|^2_{H^1(D)} \leq \frac{1}{4}\mu_0^{-1} \|f\|^2_{L^2(D)}$, whence the bound on $|u|_{H^1(D)}$. □

Remark 61.2 (A priori H^1-estimate). One can also bound the right-hand side of (61.6) by $\|f\|_{L^2(D)} C_{\mathrm{PS}}^{-1} \ell_D |u|_{H^1(D)}$, where C_{PS} comes from the Poincaré–Steklov inequality in $H^1_0(D)$ and ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. This yields

$$|u|_{H^1(D)} \leq (C_{\mathrm{PS}} \ell_D^{-1} \epsilon)^{-1} \|f\|_{L^2(D)}.$$

This bound on $|u|_{H^1(D)}$ is sharper than that in (61.5) only if $\epsilon \geq 4\mu_0 \ell_D^2 C_{\mathrm{PS}}^{-2}$. Otherwise, (61.5) is sharper and this means that the H^1-stability of the solution essentially hinges on the first-order operator A and not on the diffusion operator. In this situation, $|u|_{H^1(D)}$ behaves like $\mathcal{O}(\epsilon^{-\frac{1}{2}})$, indicating that the value of the solution can have $\mathcal{O}(1)$ variations in a layer of width ϵ. We refer the reader to Exercise 61.1 for a tighter bound on $|u|_{H^1(D)}$ and a bound on $\|\Delta u\|_{L^2(D)}$ under some more specific assumptions. □

Example 61.3 (Boundary layer). Consider the interval $D := (0,1)$ and the PDE $-\epsilon u'' + u' = f$ in D with $f := 1$, and the homogeneous Dirichlet conditions $u(0) = u(1) = 0$. One can verify that the solution is $u(x) = \left(x - \frac{e^{x/\epsilon}-1}{e^{1/\epsilon}-1}\right)$. The graph of the solution is shown in Figure 61.1 for $\epsilon \in \{1, 10^{-1}, 10^{-2}\}$. When $\epsilon \ll 1$, the solution is very close to $u_0(x) := x$ in the interval $(0, 1-\epsilon)$ (u_0 is the solution of the first-order problem $u_0' = 1$ in D and $u_0(0) = 0$), and swiftly decreases in the interval $(1-\epsilon, 1)$ to match the prescribed value $u(1) = 0$. The interval $(1-\epsilon, 1)$, $\epsilon \ll 1$, is called boundary layer (or outflow layer). □

61.2 Discrete setting

Our aim in this section is to approximate the model problem (61.2) using a shape-regular mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ and H^1-conforming finite elements of degree $k \geq 1$. To avoid technicalities, we

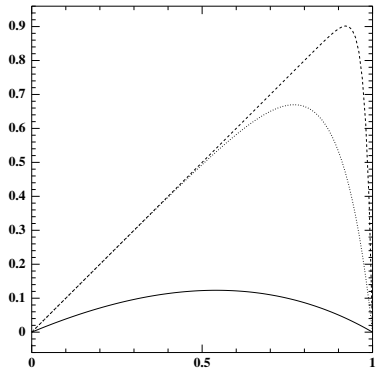

Figure 61.1: One-dimensional advection-diffusion problem with boundary layer; $-\epsilon u'' + u' = 1$ with $\epsilon \in \{1, 10^{-1}, 10^{-2}\}$.

assume that $\mathbb{d}_\epsilon$ is piecewise constant on a given partition of D and that the meshes are compatible with this partition. Hence, $\mathbb{d}_\epsilon$ is piecewise constant on $\mathcal{T}_h$ for all $h \in \mathcal{H}$, and we denote by $\lambda_{\flat,K}$ and $\lambda_{\sharp,K}$ the smallest and the largest eigenvalues of $\mathbb{d}_{\epsilon|K}$ for all $K \in \mathcal{T}_h$, respectively. The local *anisotropy ratio* is defined by $\rho_K := \lambda_{\sharp,K}/\lambda_{\flat,K}$.

We consider the local mesh-dependent weights

$$\tau_K := \min(\beta_K^{-1} h_K, \mu_0^{-1}), \tag{61.7}$$

with $\beta_K := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(K)}$ for all $K \in \mathcal{T}_h$ (see §57.3.1). Notice that τ_K represents a local time scale. The local (nondimensional) *Péclet numbers*

$$\mathrm{Pe}_K := \frac{h_K^2}{\tau_K \lambda_{\flat,K}}, \qquad \forall K \in \mathcal{T}_h, \tag{61.8}$$

are of crucial importance in the finite element approximation. One recovers the usual definition $\mathrm{Pe}_K := \frac{\beta_K h_K}{\lambda_K}$ if $\tau_K = \beta_K^{-1} h_K$ and $\mathbb{d}_\epsilon$ is isotropic (i.e., $\lambda_{\sharp,K} = \lambda_{\flat,K} =: \lambda_K$). When the mesh is fine enough, the local Péclet numbers are small, and the standard Galerkin approximation can be used to approximate the solution satisfactorily. However, it can happen that the parameter ϵ is so small that it requires very fine meshes to have small Péclet numbers. When the local Péclet numbers are large, using the standard Galerkin approximation generally leads to unacceptable discrete solutions that are globally plagued by spurious oscillations (see Exercise 61.2). In this situation, one effective remedy is to use one of the stabilized finite element methods described in the previous chapters. For brevity, we focus on the GaLS stabilization. In addition, we are going to enforce the Dirichlet boundary condition weakly by means of the boundary penalty method. This choice is motivated by the possible presence of boundary layers, where the solution is poorly approximated by discrete functions vanishing at the boundary (see Example 61.3).

Let us set $V_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$. The discrete problem is the following:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_{\epsilon h}(u_h, w_h) = \ell_{\epsilon h}(w_h), \quad \forall w_h \in V_h, \end{cases} \tag{61.9}$$

with the bilinear form $a_{\epsilon h}$ defined on $V_h \times V_h$ as follows:

$$a_{\epsilon h}(v_h, w_h) := a_\epsilon(v_h, w_h) + r_{\epsilon h}(v_h, w_h) + n_{\epsilon h}(v_h, w_h) + n_{\beta h}(v_h, w_h), \tag{61.10}$$

where

$$r_{\epsilon h}(v_h, w_h) := \sum_{K\in\mathcal{T}_h} \tau_K \delta_K (T_{\epsilon h}(v_h), T_{\epsilon h}(w_h))_{L^2(K)}, \tag{61.11a}$$

$$n_{\epsilon h}(v_h, w_h) := \sum_{F\in\mathcal{F}_h^\partial} (-\boldsymbol{n}\cdot\mathbb{d}_\epsilon \nabla v_h, w_h)_{L^2(F)} + \varpi_0 \frac{\lambda_F}{h_F}(v_h, w_h)_{L^2(F)}, \tag{61.11b}$$

$$n_{\beta h}(v_h, w_h) := \sum_{F\in\mathcal{F}_h^\partial} \frac{1}{2}\big((|\boldsymbol{\beta}\cdot\boldsymbol{n}| - \boldsymbol{\beta}\cdot\boldsymbol{n})v_h, w_h\big)_{L^2(F)}, \tag{61.11c}$$

$$\ell_{\epsilon h}(w_h) := \ell(w_h) + \sum_{K\in\mathcal{T}_h} \tau_K \delta_K (f, T_{\epsilon h}(w_h))_{L^2(K)}. \tag{61.11d}$$

The bilinear form $r_{\epsilon h}$ is the GaLS stabilization and uses the broken differential operator

$$T_{\epsilon h}(v_h) := -\nabla_h\cdot(\mathbb{d}_\epsilon \nabla v_h) + A(v_h), \tag{61.12}$$

where the notation $\nabla_h\cdot$ means that the divergence is evaluated locally in each mesh cell $K \in \mathcal{T}_h$. The blending parameter δ_K in (61.11a) is defined as

$$\delta_K := \delta(\rho_K^{-1}\mathrm{Pe}_K), \qquad \delta(\xi) := \min(1, \xi),\ \forall \xi \in \mathbb{R}_+. \tag{61.13}$$

The bilinear forms $n_{\epsilon h}$ and $n_{\beta h}$ weakly enforce the Dirichlet condition. The bilinear form $n_{\epsilon h}$ is built essentially as in Chapter 37, with the exception that we now account for the possible anisotropy in $\mathbb{d}_\epsilon$. In the penalty factor $\varpi_0\frac{\lambda_F}{h_F}$, $\varpi_0 > 0$ is a user-defined parameter to be chosen large enough (see Lemma 61.8 below), and for every boundary face $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$, we set $\lambda_F := \boldsymbol{n}\cdot(\mathbb{d}_{\epsilon|K_l}\boldsymbol{n})$ (notice that $\lambda_{\flat,K_l} \le \lambda_F \le \lambda_{\sharp,K_l}$). The bilinear form $n_{\beta h}$ is needed in the large-Péclet regime to enforce weakly the inflow boundary condition $u = 0$ on $\partial D^- := \{\boldsymbol{x} \in \partial D \mid (\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) < 0\}$; see §57.4.2 and Example 57.17 where $\mathcal{M} = |\boldsymbol{\beta}\cdot\boldsymbol{n}|$ and $\mathcal{N} = \boldsymbol{\beta}\cdot\boldsymbol{n}$ so that $\frac{1}{2}(\mathcal{M}-\mathcal{N}) = \frac{1}{2}(|\boldsymbol{\beta}\cdot\boldsymbol{n}|-\boldsymbol{\beta}\cdot\boldsymbol{n})$. Finally, the additions to the linear form ℓ in the definition of the discrete form $\ell_{\epsilon h}$ are introduced for consistency reasons.

Remark 61.4 (Parameter δ_K and function δ). The parameter δ_K ensures a smooth transition between the large-Péclet regime (where $\tau_K\delta_K = \tau_K$ scales linearly w.r.t. the meshsize and mimicks the GaLS stabilization for the first-order PDE $A(v) = f$ as in §57.3), and the small-Péclet regime (where $\tau_K\delta_K$ decays quadratically w.r.t. the meshsize). The use of the anisotropy factor ρ_K^{-1} in the estimation of δ_K is motivated by the error analysis. Several choices are actually possible for the function δ in (61.13) provided one has $c_1\min(1, x) \le \delta(x) \le c_2\min(1, x)$ for some constants c_1, c_2. For instance, one can use the function $\delta(x) := \coth(\frac{x}{2}) - \frac{2}{x}$, (sometimes called *Scharfetter–Gummel function* in the literature). □

Remark 61.5 (Literature). The finite element approximation of advection-diffusion equations is covered in many textbooks as, e.g., Quarteroni and Valli [239, p. 269], Roos et al. [243, p. 277]. All the stabilization methods from the previous chapters can be used to approximate *singularly perturbed* first-order PDEs. We refer the reader to Burman [57], Burman and Hansbo [67] for CIP, Guermond [144] for SGV, and Braack and Burman [41], Matthies et al. [227] for LPS. Concerning dG methods, we mention Houston et al. [186] for the *hp*-analysis, and Di Pietro et al. [106], Ern et al. [124] for weighted averages and harmonic penalties (see also Di Pietro and Ern [105, §4.6]). The weak enforcement of boundary conditions in the advection-dominated (large-Péclet) regime has been motivated numerically in Bazilevs and Hughes [27] and analyzed in Schieweck [246]; see also Burman et al. [72]. □

Remark 61.6 (Dominant reaction). The present GaLS stabilization can also be used in the dominant reaction regime, e.g., $\beta_K \ll \mu_0 h_K$. In this case, $\tau_K = \mu_0^{-1}$, and the local Péclet number becomes $\mathrm{Pe}_K = \frac{h_K^2 \mu_0}{\lambda_{\flat,K}}$, i.e., Pe_K scales quadratically with the meshsize. □

61.3 Stability and error analysis

This section is devoted to the stability and error analysis of the discrete problem (61.9).

61.3.1 Stability and well-posedness

We equip the discrete space V_h with the norm

$$\|v_h\|_{V_h}^2 := \|\mathrm{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + \mu_0\|v_h\|_{L^2(D)}^2 + \sum_{K\in\mathcal{T}_h}\tau_K\delta_K\|T_{\epsilon h}(v_h)\|_{L^2(K)}^2 + \sum_{F\in\mathcal{F}_h^\partial}\frac{\lambda_F}{h_F}\|v_h\|_{L^2(F)}^2 + \frac12\sum_{F\in\mathcal{F}_h^\partial}\||\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac12}v_h\|_{L^2(F)}^2. \tag{61.14}$$

As in §37.2, we consider the smallest constant c_{dt} such that the discrete trace inequality $\|v_h\|_{L^2(F)} \le c_{\mathrm{dt}}h_F^{-\frac12}\|v_h\|_{L^2(K_l)}$ holds for all $v_h \in V_h$ and all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. We denote by n_∂ the maximum number of boundary faces a cell K_l can have ($n_\partial \le d$ for simplicial meshes). We start with a bound on the consistency term associated with the diffusion part of the boundary penalty bilinear form.

Lemma 61.7 (Bound on consistency term). *Let $\mathcal{T}_h^{\partial D}$ be the collection of the mesh cells having at least one boundary face, i.e., $\mathcal{T}_h^{\partial D} := \bigcup_{F\in\mathcal{F}_h^\partial}\{K_l\}$. The following holds true for all $v_h, w_h \in V_h$:*

$$\left|\int_{\partial D}(\boldsymbol{n}\cdot\mathrm{d}_\epsilon\nabla v_h)w_h\,\mathrm{d}s\right| \le n_\partial^{\frac12}c_{\mathrm{dt}}\Big(\sum_{K\in\mathcal{T}_h^{\partial D}}\|\mathrm{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(K)}^2\Big)^{\frac12}\Big(\sum_{F\in\mathcal{F}_h^\partial}\frac{\lambda_F}{h_F}\|w_h\|_{L^2(F)}^2\Big)^{\frac12}. \tag{61.15}$$

Proof. See Exercise 61.4. □

Lemma 61.8 (Coercivity, well-posedness). *Assume that the penalty parameter is such that $\varpi_0 \ge 1 + \frac12 n_\partial c_{\mathrm{dt}}^2$.* (i) *The following holds true:*

$$a_{\epsilon h}(v_h, v_h) \ge \frac12\|v_h\|_{V_h}^2, \qquad \forall v_h \in V_h. \tag{61.16}$$

(ii) *The discrete problem* (61.9) *is well-posed.*

Proof. We only need to prove (61.16) since the well-posedness of (61.9) then follows from the Lax–Milgram lemma. Rearranging the terms, we observe that for all $v_h \in V_h$,

$$\begin{aligned} a_{\epsilon h}(v_h, v_h) &= \Big((\mathrm{d}_\epsilon\nabla v_h, \nabla v_h)_{\boldsymbol{L}^2(D)} + n_{\epsilon h}(v_h, v_h)\Big) \\ &\quad + \Big((A(v_h), v_h)_{L^2(D)} + n_{\beta h}(v_h, v_h)\Big) \\ &\quad + r_{\epsilon h}(v_h, v_h) =: \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3. \end{aligned}$$

Owing to Lemma 61.7, we infer that $\mathfrak{T}_1 \geq z^2 + (x^2 - n_\partial^{\frac12} c_{\mathrm{dt}} xy + \varpi_0 y^2)$ with

$$z := \Big(\sum_{K\in\mathcal{T}_h\setminus\mathcal{T}_h^{\partial D}} \|\mathrm{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(K)}^2 \Big)^{\frac12}, \qquad x := \Big(\sum_{K\in\mathcal{T}_h^{\partial D}} \|\mathrm{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(K)}^2 \Big)^{\frac12},$$
$$y := \Big(\sum_{F\in\mathcal{F}_h^\partial} \frac{\lambda_F}{h_F}\|v_h\|_{L^2(F)}^2 \Big)^{\frac12}.$$

Using the quadratic inequality invoked in the proof of Lemma 37.3 (i.e., $x^2 - n_\partial^{\frac12} c_{\mathrm{dt}} xy + \varpi_0 y^2 \geq \frac{\varpi_0 - \frac14 n_\partial c_{\mathrm{dt}}^2}{1+\varpi_0}(x^2+y^2)$) and since $\frac{\varpi_0 - \frac14 n_\partial c_{\mathrm{dt}}^2}{1+\varpi_0} \geq \frac12$ (because we assumed that $\varpi_0 \geq 1 + \frac12 n_\partial c_{\mathrm{dt}}^2$), we obtain

$$\mathfrak{T}_1 \geq \frac12 \Big(\|\mathrm{d}_\epsilon^{\frac12}\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + \sum_{F\in\mathcal{F}_h^\partial} \frac{\lambda_F}{h_F}\|v_h\|_{L^2(F)}^2 \Big).$$

Furthermore, proceeding as in Lemma 57.20, we infer that

$$\mathfrak{T}_2 \geq \mu_0 \|v_h\|_{L^2(D)}^2 + \frac12 \sum_{F\in\mathcal{F}_h^\partial} \||\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac12} v_h\|_{L^2(F)}^2,$$

and we have $\mathfrak{T}_3 := \sum_{K\in\mathcal{T}_h} \tau_K \delta_K \|T_{\epsilon h}(v_h)\|_{L^2(K)}^2$. Collecting the above estimates shows that (61.16) holds true. □

Remark 61.9 (Penalty parameter). Any value $\varpi_0 > \frac14 n_\partial c_{\mathrm{dt}}^2$ yields coercivity, and the present choice allows us to make the coercivity constant equal to $\frac12$. Let us also mention that the proof of Lemma 61.7 uses the fact that d_ϵ is piecewise constant, making $\mathrm{d}_\epsilon\nabla v_h$ a piecewise polynomial function. In the more general situation where d_ϵ is piecewise smooth (e.g., Lipschitz) in each mesh cell, the scaling of the penalty term in (61.11b) should be $\varpi_0 \rho_{K_l} \frac{\lambda_F}{h_F}$. □

61.3.2 Consistency/boundedness

We are going to use the setting of Lemma 27.8 to perform the error analysis. Let u be the solution to the model problem (61.2). We assume that

$$u \in V_{\mathrm{s}} := H_0^1(D) \cap H^{1+r}(D), \qquad r \geq 1. \tag{61.17}$$

We set $V_\sharp := V_{\mathrm{s}} + V_h$, and we equip this space with the following two norms:

$$\|v\|_{V_\flat}^2 := \|\mathrm{d}_\epsilon^{\frac12}\nabla v\|_{\boldsymbol{L}^2(D)}^2 + \mu_0\|v\|_{L^2(D)}^2 + \sum_{K\in\mathcal{T}_h} \tau_K\delta_K\|T_{\epsilon h}(v)\|_{L^2(K)}^2$$
$$+ \sum_{F\in\mathcal{F}_h^\partial} \frac{\lambda_F}{h_F}\|v\|_{L^2(F)}^2 + \frac12 \sum_{F\in\mathcal{F}_h^\partial} \||\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac12} v\|_{L^2(F)}^2, \tag{61.18a}$$

$$\|v\|_{V_\sharp}^2 := \|v\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h} \rho_K^{\frac12}\tau_K^{-1}\|v\|_{L^2(K)}^2 + \sum_{F\in\mathcal{F}_h^\partial} h_F \|\mathrm{d}_\epsilon^{\frac12}\nabla v\|_{\boldsymbol{L}^2(F)}^2. \tag{61.18b}$$

We observe that (27.7) holds true with $c_\flat := 1$ (i.e., $\|v_h\|_{V_\flat} \leq \|v_h\|_{V_h}$ on V_h and $\|v\|_{V_\flat} \leq \|v\|_{V_\sharp}$ on $V_\sharp$). To simplify the tracking of model-dependent constants, we assume (as in Chapter 57) that $\max(\|\mu\|_{L^\infty(D)}, \|\nabla\cdot\boldsymbol{\beta}\|_{L^\infty(D)}) \leq c_{\mu,\beta}\mu_0$, and we hide the quantity $c_{\mu,\beta}$ in the generic constants appearing in the error analysis.

Lemma 61.10 (Consistency/boundedness). *Define the consistency error as*

$$\langle \delta_h(v_h), w_h\rangle_{V_h',V_h} := \ell_h(w_h) - a_{\epsilon h}(v_h,w_h), \qquad \forall v_h,w_h\in V_h.$$

There is $\omega_\sharp$, uniform w.r.t. $u\in V_{\rm S}$, s.t. for all $v_h,w_h\in V_h$, all $h\in\mathcal{H}$, and all d_ϵ,

$$|\langle \delta_h(v_h), w_h\rangle_{V_h',V_h}| \le \omega_\sharp \|u-v_h\|_{V_\sharp}\|w_h\|_{V_h}. \tag{61.19}$$

Proof. (1) Since $u\in H^{1+r}(D)$ with $r\ge 1$, we infer that $\boldsymbol{n}\cdot\mathrm{d}_\epsilon u$ has a well-defined trace on ∂D and that $T_\epsilon(u) = T_{\epsilon h}(u) = f\in L^2(D)$. Hence, we can write $a_\epsilon(u,w_h) = (T_\epsilon(u),w_h)_{L^2(D)} + (\boldsymbol{n}\cdot\mathrm{d}_\epsilon u, w_h)_{L^2(\partial D)}$ for all $w_h\in V_h$. Since u vanishes at ∂D, we obtain

$$\ell_{\epsilon h}(w_h) = a_\epsilon(u,w_h) + n_{\epsilon h}(u,w_h) + n_{\beta h}(u,w_h) + r_{\epsilon h}(u,w_h).$$

Putting these identities together, we infer that $\langle \delta_h(v_h),w_h\rangle_{V_h',V_h} = \mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3$ with $\mathfrak{T}_1 := (\mathrm{d}_\epsilon\nabla\eta,\nabla w_h)_{\boldsymbol{L}^2(D)} + n_{\epsilon h}(\eta,w_h)$, $\mathfrak{T}_2 := (A(\eta),w_h)_{L^2(D)} + n_{\beta h}(\eta,w_h)$, $\mathfrak{T}_3 := r_{\epsilon h}(\eta,w_h)$, and $\eta := u - v_h$.

(2) The term $\mathfrak{T}_1$ is estimated as in the proof of Lemma 37.5, where we now use that d_ϵ is symmetric positive definite to write $|\boldsymbol{n}\cdot\mathrm{d}_\epsilon\nabla\eta| \le \lambda_F^{\frac12}\|\mathrm{d}_\epsilon^{\frac12}\nabla\eta\|_{\ell^2(\mathbb{R}^d)}$, and invoking the Cauchy–Schwarz inequality, we infer that

$$\big|(\boldsymbol{n}\cdot\mathrm{d}_\epsilon\nabla\eta,w_h)_{L^2(\partial D)}\big| \le \Big(\sum_{F\in\mathcal{F}_h^\partial} h_F\|\mathrm{d}_\epsilon^{\frac12}\nabla\eta\|_{\boldsymbol{L}^2(F)}^2\Big)^{\frac12}\Big(\sum_{F\in\mathcal{F}_h^\partial}\frac{\lambda_F}{h_F}\|w_h\|_{L^2(F)}^2\Big)^{\frac12}.$$

The term $\mathfrak{T}_3$ is bounded by using the Cauchy–Schwarz inequality. We proceed almost exactly as in the proof of Lemma 57.21 to estimate the term $\mathfrak{T}_2$, the only difference being the bound on the term $\mathfrak{T}_2' := -(\eta,\boldsymbol{\beta}\cdot\nabla w_h)_{L^2(D)}$. The Cauchy–Schwarz inequality yields

$$|\mathfrak{T}_2'| \le \sum_{K\in\mathcal{T}_h}\|\eta\|_{L^2(K)}\|\boldsymbol{\beta}\cdot\nabla w_h\|_{L^2(K)},$$

and we distinguish two cases.

(2a) Assume that $\mathrm{Pe}_K\le\rho_K^{\frac12}$. Then we have

$$\begin{aligned}\|\eta\|_{L^2(K)}\|\boldsymbol{\beta}\cdot\nabla w_h\|_{L^2(K)} &\le \|\eta\|_{L^2(K)}\beta_K\lambda_{\flat,K}^{-\frac12}\|\mathrm{d}_\epsilon^{\frac12}\nabla w_h\|_{\boldsymbol{L}^2(K)}\\ &\le \rho_K^{\frac14}\tau_K^{-\frac12}\|\eta\|_{L^2(K)}\|\mathrm{d}_\epsilon^{\frac12}\nabla w_h\|_{\boldsymbol{L}^2(K)},\end{aligned}$$

since $\beta_K\lambda_{\flat,K}^{-\frac12} = \beta_K(\mathrm{Pe}_K\tau_K h_K^{-2})^{\frac12} \le \rho_K^{\frac14}\beta_K h_K^{-1}\tau_K^{\frac12} \le \rho_K^{\frac14}\tau_K^{-1}\tau_K^{\frac12} = \rho_K^{\frac14}\tau_K^{-\frac12}$.

(2b) Assume now that $\mathrm{Pe}_K\ge\rho_K^{\frac12}$. Up to the zero-order term μw_h which is bounded as usual, we use the triangle inequality and obtain

$$\begin{aligned}\|\eta\|_{L^2(K)}\|A(w_h)\|_{L^2(K)} &\le \|\eta\|_{L^2(K)}(\|T_{\epsilon h}(w_h)\|_{L^2(K)} + \|\nabla\cdot(\mathrm{d}_\epsilon\nabla w_h)\|_{L^2(K)})\\ &\le \tau_K^{-\frac12}\|\eta\|_{L^2(K)}(\tau_K^{\frac12}\|T_{\epsilon h}(w_h)\|_{L^2(K)} + c\,\tau_K^{\frac12}h_K^{-1}\lambda_{\sharp,K}^{\frac12}\|\mathrm{d}_\epsilon^{\frac12}\nabla w_h\|_{\boldsymbol{L}^2(K)}),\end{aligned}$$

where we used an inverse inequality. Since $\mathrm{Pe}_K\ge\rho_K^{\frac12}$ by assumption, we have $\tau_K^{\frac12}h_K^{-1}\lambda_{\sharp,K}^{\frac12} = \mathrm{Pe}_K^{-\frac12}\rho_K^{\frac12} \le \rho_K^{\frac14}$, and since the function δ is nondecreasing and satisfies $\delta(z)\le z$, we also have $\delta_K^{-\frac12} := \big(\delta(\rho_K^{-1}\mathrm{Pe}_K)\big)^{-\frac12} \le \big(\delta(\rho_K^{-\frac12})\big)^{-\frac12} \le \rho_K^{\frac14}$, i.e., $\rho_K^{-\frac14}\le\delta_K^{\frac12}$. We finally infer that

$$\|\eta\|_{L^2(K)}\|A(w_h)\|_{L^2(K)} \le c\,\rho_K^{\frac14}\tau_K^{-\frac12}\|\eta\|_{L^2(K)}\times\Big((\tau_K\delta_K)^{\frac12}\|T_{\epsilon h}(w_h)\|_{L^2(K)} + \|\mathrm{d}_\epsilon^{\frac12}\nabla w_h\|_{\boldsymbol{L}^2(K)}\Big).$$

Collecting the above bounds leads to the expected estimate. □

61.3.3 Error estimates

Recall that D_K is the set of the points composing the mesh cells sharing at least a vertex with the mesh cell $K \in \mathcal{T}_h$.

Theorem 61.11 (Error estimate). *Let u solve (61.2) and assume $u \in V_\mathrm{s}$. Assume that the penalty parameter ϖ_0 is s.t. $\varpi_0 \geq 1 + \frac{1}{2} n_\partial c_{\mathrm{dt}}^2$.* (i) *There is c such that for all $h \in \mathcal{H}$ and $\mathbb{d}_\epsilon$,*

$$\|u - u_h\|_{V_\flat} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{61.20}$$

(ii) *Provided $r \in [1, k]$, then*

$$\|u - u_h\|_{V_\flat} \leq c \left(\sum_{K \in \mathcal{T}_h} \left(\rho_K^{\frac{1}{2}} (\mu_0 h_K^2 + \beta_K h_K) + \lambda_{\sharp,K} \right) h_K^{2r} |u|^2_{H^{1+r}(D_K)} \right)^{\frac{1}{2}}, \tag{61.21}$$

and $|u|_{H^{1+r}(D_K)}$ can be replaced by $|u|_{H^{1+r}(K)}$ if $1 + r > \frac{d}{2}$.

Proof. (i) The estimate (61.20) follows from Lemma 27.8 combined with Lemma 61.8 (stability) and Lemma 61.10 (consistency/boundedness).
(ii) We pick $v_h := \mathcal{I}_h(u)$ to prove (61.21), where the quasi-interpolation operator $\mathcal{I}_h : L^1(D) \to V_h$ satisfies the following local optimal approximation property (see Theorem 22.6) for all $m \in \{0, 1, 2\}$ (taking $m := 2$ is allowed since $r \geq 1$):

$$|u - \mathcal{I}_h(u)|_{H^m(K)} \leq c\, h_K^{1+r-m} |u|_{H^{1+r}(D_K)}. \tag{61.22}$$

We have to estimate $\|u - \mathcal{I}_h(u)\|_{V_\sharp}$. Among the terms composing $\|\cdot\|_{V_\sharp}$, we only bound

$$\sum_{K \in \mathcal{T}_h} \rho_K^{\frac{1}{2}} \tau_K^{-1} \|u - \mathcal{I}_h(u)\|^2_{L^2(K)} \quad \text{and} \quad \sum_{K \in \mathcal{T}_h} \tau_K \delta_K \|T_{\epsilon h}(u - \mathcal{I}_h(u))\|^2_{L^2(K)},$$

since the others can be estimated as in the previous chapters. Since $\tau_K^{-1} \leq \mu_0 + \beta_K h_K^{-1}$, using (61.22) with $m := 0$ gives $\rho_K^{\frac{1}{2}} \tau_K^{-1} \|u - \mathcal{I}_h(u)\|^2_{L^2(K)} \leq c \rho_K^{\frac{1}{2}} (\mu_0 h_K^2 + \beta_K h_K) h_K^{2r} |u|^2_{H^{1+r}(D_K)}$. Let us now derive a bound on $\tau_K \delta_K \|T_{\epsilon h}(u - \mathcal{I}_h(u))\|^2_{L^2(K)}$. The triangle inequality yields

$$\|T_{\epsilon h}(u - \mathcal{I}_h(u))\|_{L^2(K)} \leq \|A(u - \mathcal{I}_h(u))\|_{L^2(K)} + \|\nabla_h \cdot (\mathbb{d}_\epsilon \nabla (u - \mathcal{I}_h(u)))\|_{L^2(K)}.$$

Using (61.22) with $m \in \{0, 1\}$, we infer that

$$\|A(u - \mathcal{I}_h(u))\|_{L^2(K)} \leq c\, (\mu_0 h_K + \beta_K) h_K^r |u|_{H^{1+r}(D_K)}.$$

Moreover, since $\mathbb{d}_\epsilon$ is constant on K, using (61.22) with $m := 2$ yields

$$\begin{aligned} \|\nabla_h \cdot (\mathbb{d}_\epsilon \nabla (u - \mathcal{I}_h(u)))\|_{L^2(K)} &\leq c\, \lambda_{\sharp,K} |u - \mathcal{I}_h(u)|_{H^2(K)} \\ &\leq c\, h_K^{-1} \lambda_{\sharp,K} h_K^r |u|_{H^{1+r}(D_K)}. \end{aligned}$$

Since $\delta_K := \min(1, \rho_K^{-1} \mathrm{Pe}_K)$ and $\tau_K := \min(\beta_K^{-1} h_K, \mu_0^{-1})$, we infer that $\tau_K^{\frac{1}{2}} \delta_K^{\frac{1}{2}} (\mu_0 h_K + \beta_K) \leq \tau_K^{\frac{1}{2}} (\mu_0 h_K + \beta_K) \leq (\mu_0^{\frac{1}{2}} h_K + \beta_K^{\frac{1}{2}} h_K^{\frac{1}{2}})$ and $\tau_K^{\frac{1}{2}} \delta_K^{\frac{1}{2}} h_K^{-1} \lambda_{\sharp,K} \leq \tau_K^{\frac{1}{2}} \rho_K^{-\frac{1}{2}} \mathrm{Pe}_K^{\frac{1}{2}} h_K^{-1} \lambda_{\sharp,K} = \lambda_{\sharp,K}^{\frac{1}{2}}$. Hence, we have

$$\tau_K^{\frac{1}{2}} \delta_K^{\frac{1}{2}} \|T_{\epsilon h}(u - \mathcal{I}_h(u))\|_{L^2(K)} \leq c\, (\mu_0^{\frac{1}{2}} h_K + \beta_K^{\frac{1}{2}} h_K^{\frac{1}{2}} + \lambda_{\sharp,K}^{\frac{1}{2}}) h_K^r |u|_{H^{1+r}(D_K)}.$$

We conclude the proof of (61.21) by using that $\rho_K \geq 1$. □

Corollary 61.12 (Asymptotic regimes). *Let the assumptions of Theorem* 61.11 *hold true.*

(i) *If* $\mathrm{Pe}_K \le \rho_K^{\frac12}$ *for all* $K \in \mathcal{T}_h$ *(dominant diffusion), we have*

$$\|\mathrm{d}_\epsilon^{\frac12}\nabla(u-u_h)\|_{\boldsymbol{L}^2(D)} \le c\left(\sum_{K\in\mathcal{T}_h}\lambda_{\sharp,K}h_K^{2r}|u|^2_{H^{1+r}(D_K)}\right)^{\frac12}. \tag{61.23}$$

(ii) *If* $\mathrm{Pe}_K \ge \rho_K^{\frac12}$ *for all* $K \in \mathcal{T}_h$ *(dominant advection), we have*

$$\mu_0^{\frac12}\|u-u_h\|_{L^2(D)} + \left(\sum_{K\in\mathcal{T}_h}\rho_K^{-\frac12}\tau_K\|A(u-u_h)\|^2_{L^2(K)}\right)^{\frac12} \le c\left(\sum_{K\in\mathcal{T}_h}\rho_K^{\frac12}(\mu_0h_K+\beta_K)h_K^{2r+1}|u|^2_{H^{1+r}(D_K)}\right)^{\frac12}. \tag{61.24}$$

Proof. We have $\tau_K^{-1}h_K^2 \le \mu_0h_K^2+\beta_Kh_K \le 2\tau_K^{-1}h_K^2$ since $\tau_K := \min(\frac{h_K}{\beta_K},\frac{1}{\mu_0})$ and $\frac{1}{\min(a,b)} \le \frac1a+\frac1b \le \frac{2}{\min(a,b)}$ for any positive real numbers a,b.

(i) Assume that $\mathrm{Pe}_K \le \rho_K^{\frac12}$ for all $K\in\mathcal{T}_h$. Then we have $\rho_K^{\frac12}(\mu_0h_K^2+\beta_Kh_K) \le 2\rho_K^{\frac12}\tau_K^{-1}h_K^2 = 2\lambda_{\flat,K}\mathrm{Pe}_K\rho_K^{\frac12} \le 2\lambda_{\sharp,K}$. Therefore, the bound (61.21) becomes

$$\|u-u_h\|_{V_\flat} \le c\left(\sum_{K\in\mathcal{T}_h}\lambda_{\sharp,K}h_K^{2r}|u|^2_{H^{1+r}(D_K)}\right)^{\frac12},$$

and (61.23) follows since $\|\mathrm{d}_\epsilon^{\frac12}\nabla(u-u_h)\|_{\boldsymbol{L}^2(D)} \le \|u-u_h\|_{V_\flat}$.

(ii) Assume that $\mathrm{Pe}_K \ge \rho_K^{\frac12}$ for all $K\in\mathcal{T}_h$. We infer that $\lambda_{\sharp,K} = \rho_K^{\frac12}\tau_K^{-1}h_K^2\frac{\rho_K^{\frac12}}{\mathrm{Pe}_K} \le \rho_K^{\frac12}\tau_K^{-1}h_K^2 \le \rho_K^{\frac12}(\mu_0h_K^2+\beta_Kh_K)$. Therefore, the bound (61.21) becomes $\|u-u_h\|_{V_\flat} \le c(\sum_{K\in\mathcal{T}_h}\rho_K^{\frac12}(\mu_0h_K^2+\beta_Kh_K)h_K^{2r}|u|^2_{H^{1+r}(D_K)})^{\frac12}$. Since $\mu_0^{\frac12}\|u-u_h\|_{L^2(D)} \le \|u-u_h\|_{V_\flat}$, this proves the estimate on $\mu_0^{\frac12}\|u-u_h\|_{L^2(D)}$ in (61.24). It remains to estimate $(\sum_{K\in\mathcal{T}_h}\rho_K^{-\frac12}\tau_K\|A(u-u_h)\|^2_{L^2(K)})^{\frac12}$. Using the triangle inequality and the inequality $\rho_K^{-\frac14} \le \delta_K^{\frac12}$ (see Step (2b) of the proof of Lemma 61.10), we infer that

$$\rho_K^{-\frac14}\tau_K^{\frac12}\|A(u-u_h)\|_{L^2(K)} \le \tau_K^{\frac12}\delta_K^{\frac12}\|T_{\epsilon h}(u-u_h)\|_{L^2(K)} + \rho_K^{-\frac14}\tau_K^{\frac12}\|\nabla\cdot(\mathrm{d}_\epsilon\nabla(u-u_h))\|_{L^2(K)}.$$

Let us consider the second term on the right-hand side. Using the approximation properties of the operator $\mathcal{I}_h$, an inverse inequality, and the triangle inequality, we obtain

$$\begin{aligned}\|\nabla\cdot(\mathrm{d}_\epsilon\nabla(u-u_h))\|_{L^2(K)} &\le \|\nabla\cdot(\mathrm{d}_\epsilon\nabla(u-\mathcal{I}_hu))\|_{L^2(K)} + \|\nabla\cdot(\mathrm{d}_\epsilon\nabla(\mathcal{I}_hu-u_h))\|_{L^2(K)}\\ &\le c\left(\lambda_{\sharp,K}h_K^{r-1}|u|_{H^{1+r}(D_K)} + h_K^{-1}\lambda_{\sharp,K}^{\frac12}\|\mathrm{d}_\epsilon^{\frac12}\nabla(\mathcal{I}_hu-u_h)\|_{\boldsymbol{L}^2(K)}\right)\\ &\le c'\left(h_K^{-1}\lambda_{\sharp,K}h_K^r|u|_{H^{1+r}(D_K)} + h_K^{-1}\lambda_{\sharp,K}^{\frac12}\|\mathrm{d}_\epsilon^{\frac12}\nabla(u-u_h)\|_{\boldsymbol{L}^2(K)}\right).\end{aligned}$$

Observing that $\rho_K^{-\frac14}\tau_K^{\frac12} \le h_K\lambda_{\sharp,K}^{-\frac12}$ if $\mathrm{Pe}_K \ge \rho_K^{\frac12}$, we infer that

$$\rho_K^{-\frac14}\tau_K^{\frac12}\|\nabla\cdot(\mathrm{d}_\epsilon\nabla(u-u_h))\|_{L^2(K)} \le c\left(\lambda_{\sharp,K}^{\frac12}h_K^r|u|_{H^{1+r}(D_K)} + \|\mathrm{d}_\epsilon^{\frac12}\nabla(u-u_h)\|_{\boldsymbol{L}^2(K)}\right).$$

Combining these bounds and recalling the definition of $\|\cdot\|_{V_\flat}$ in (61.18a) gives

$$\sum_{K\in\mathcal{T}_h} \rho_K^{-\frac{1}{2}}\tau_K\|A(u-u_h)\|_{L^2(K)}^2 \le c\Big(\|u-u_h\|_{V_\flat}^2 + \sum_{K\in\mathcal{T}_h}\lambda_{\sharp,K}h_K^{2r}|u|_{H^{1+r}(D_K)}^2\Big).$$

Since we have already established that $\lambda_{\sharp,K} \le \rho_K^{\frac{1}{2}}(\mu_0 h_K^2 + \beta_K h_K)$, this completes the proof of (61.24). □

Remark 61.13 (Anisotropy). The dependence of the error estimate on the global anisotropy ratio $\rho := \max_{K\in\mathcal{T}_h}\rho_K$ is very mild. The error estimate in (61.21) and the bound on $\mu_0^{\frac{1}{2}}\|u-u_h\|_{L^2(D)}$ in Corollary 61.12 both scale as $\rho^{\frac{1}{4}}$. The bound on $\|\mathrm{d}_\epsilon^{\frac{1}{2}}\nabla(u-u_h)\|_{\boldsymbol{L}^2(D)}$ is robust w.r.t. ρ, and the bound on $A(u-u_h)$ scales as $\rho^{\frac{1}{2}}$. The error analysis with anisotropic diffusion is more intricate for fluctuation-based stabilization and for discontinuous Galerkin methods than it is for GaLS since stability then hinges on an inf-sup condition and not just coercivity. (The difficulty in the proof of the inf-sup condition appears when bounding $\|\mathrm{d}_\epsilon^{\frac{1}{2}}\nabla(\mathcal{J}_h(\boldsymbol{\beta}\cdot\nabla v_h))\|_{\boldsymbol{L}^2(D)}$ for all $v_h\in V_h$, where $\mathcal{J}_h$ is some averaging operator.) □

Remark 61.14 (Localization). The above convergence results feature a high-order Sobolev norm of the solution to (61.2) which can be quite large if the solution has internal or boundary layers. A refinement of the analysis for GaLS stabilization using Sobolev norms weighted by cut-off functions with exponential decay gives localized error estimates away from the layers; see Johnson et al. [201, 202]. These estimates essentially show that the GaLS-stabilized discrete solution is well-behaved away from the layers, contrary to the standard Galerkin approximation where spurious oscillations are global. Similar results have been derived for dG methods in Guzmán [173] and for CIP stabilization in Burman et al. [73]. □

61.4 Divergence-free advection

The above analysis hinges on the assumption that $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \ge \mu_0 > 0$ a.e. in D. The goal of this section is to analyze the GaLS approximation of the model problem (61.1) under the weaker assumption $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \ge 0$ a.e. in D. This setting covers in particular the case of zero reaction and divergence-free advection. Following Devinatz et al. [104], we assume that there is a function $\zeta\in C^{0,1}(D)$ such that

$$-\frac{1}{2}\boldsymbol{\beta}\cdot\nabla\zeta \ge \mu_0 > 0 \quad \text{a.e. in } D, \tag{61.25}$$

and since ζ can be defined up to any additive constant, we can assume that $\zeta \ge 1$ a.e. in D. The assumption (61.25) is reasonable whenever the field $\boldsymbol{\beta}$ has no closed streamlines and no stationary point in D. The one-dimensional version of this problem is investigated in §27.3.2 and in Exercise 27.4.

We are going to show that under the assumption (61.25), the discrete bilinear form $a_{\epsilon h}$ defined in (61.11) still enjoys stability in the norm $\|\cdot\|_{V_h}$ defined in (61.14), but this time stability follows from an inf-sup condition instead of coercivity. Once the inf-sup stability is established, the rest of the error analysis is unmodified, that is, the error estimates in Theorem 61.11 and Corollary 61.12 still hold true. Let us set $\zeta_\sharp := \|\zeta\|_{L^\infty(D)}$ and let us denote by L_ζ the Lipschitz constant of ζ in D (L_ζ scales like the reciprocal of a length). To simplify the tracking of parameter-dependent constants, we make the mild assumptions that $L_\zeta^2\max(\lambda_{\sharp,K},\beta_K h_K) \le \mu_0$ and $L_\zeta h_K \le 1$ for all $K\in\mathcal{T}_h$. The generic constants may depend on $\zeta_\sharp$ in what follows.

Lemma 61.15 (Stability). *Assume* (61.25) *and* $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \geq 0$ *a.e. in* D. *Assume the tightened stability condition* $\varpi_0 \geq 1 + \frac{1}{2}n_\partial c_{\mathrm{dt}}^2\zeta_\sharp$. *There is* $\alpha > 0$ *such that for all* $h \in \mathcal{H}$ *and* d_ϵ,

$$\alpha\,\|v_h\|_{V_h} \leq \sup_{w_h\in V_h}\frac{|a_{\epsilon h}(v_h,w_h)|}{\|w_h\|_{V_h}}, \qquad \forall v_h \in V_h. \tag{61.26}$$

Proof. We only sketch the proof. Let us set $A_1^2 := \|\mathrm{d}_\epsilon^{\frac{1}{2}}\nabla v_h\|_{\boldsymbol{L}^2(D)}^2$, $A_2^2 := \sum_{F\in\mathcal{F}_h^\partial}\lambda_F h_F^{-1}\|v_h\|_{L^2(F)}^2$, $A_3^2 := \sum_{K\in\mathcal{T}_h}\tau_K\delta_K\|T_{\epsilon h}(v_h)\|_{L^2(K)}^2$, and $A_4^2 := \frac{1}{2}\sum_{F\in\mathcal{F}_h^\partial}\||\boldsymbol{\beta}\cdot\boldsymbol{n}|^{\frac{1}{2}}v_h\|_{L^2(F)}^2$, so that $\|v_h\|_{V_h}^2 = A_1^2 + A_2^2 + A_3^2 + A_4^2 + \mu_0\|v_h\|_{L^2(D)}^2$ for all $v_h \in V_h$. Since $\varpi_0 \geq 1 + \frac{1}{2}n_\partial c_{\mathrm{dt}}^2$ (recall that $\zeta_\sharp \geq 1$) and since $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \geq 0$ a.e. in D, we infer that

$$a_{\epsilon h}(v_h,v_h) \geq \frac{1}{2}(A_1^2 + A_2^2) + A_3^2 + A_4^2.$$

Let us set $\varphi_h := \mathcal{J}_h^{\mathrm{g,av}}(\zeta_0 v_h)$, where $\mathcal{J}_h^{\mathrm{g,av}} : P_k^{\mathrm{b}}(\mathcal{T}_h) \to V_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$ is the H^1-conforming averaging operator introduced in §22.2, and $\zeta_0 := \mathcal{I}_{0,h}^{\mathrm{b}}(\zeta)$ where $\mathcal{I}_{0,h}^{\mathrm{b}} : L^2(D) \to P_0^{\mathrm{b}}(\mathcal{T}_h)$ is the L^2-orthogonal projection onto piecewise constant functions on $\mathcal{T}_h$. Let us define the bilinear form $a_1 : V_h \times P_k^{\mathrm{b}}(\mathcal{T}_h) \to \mathbb{R}$ by setting

$$a_1(v_h,w_h) := (\mathrm{d}_\epsilon\nabla v_h, \nabla_h w_h)_{\boldsymbol{L}^2(D)} + n_{\epsilon h}(v_h,w_h) + r_{\epsilon h}(v_h,w_h),$$

for all $(v_h,w_h) \in V_h\times P_k^{\mathrm{b}}(\mathcal{T}_h)$. Let us also define the bilinear form $a_2 : V_h\times H^1(\mathcal{T}_h) \to \mathbb{R}$ by setting

$$a_2(v_h,w) := (A(v_h),w)_{L^2(D)} + n_{\beta h}(v_h,w),$$

for all $(v_h,w) \in V_h\times H^1(\mathcal{T}_h)$ (recall that $H^1(\mathcal{T}_h) := \{v \in L^2(D) \mid v_{|K} \in H^1(K)\ \forall K \in \mathcal{T}_h\}$). We notice that $a_{\epsilon h} := a_{1|V_h\times V_h} + a_{2|V_h\times V_h}$. Since ζ_0 is piecewise constant, $\zeta \geq 1$, and $\varpi_0 \geq 1 + \frac{1}{2}n_\partial c_{\mathrm{dt}}^2\zeta_\sharp$, we have

$$a_1(v_h,\zeta_0 v_h) \geq \frac{1}{2}(A_1^2 + A_2^2) + A_3^2.$$

Moreover, one can show (see Exercise 61.5) that there is $c_1 > 0$ such that

$$a_1(v_h,\varphi_h - \zeta_0 v_h) \geq -c_1\varpi_0^{\frac{1}{2}}(A_1 + A_2 + A_3)\times\mu_0^{\frac{1}{2}}\|v_h\|_{L^2(D)}. \tag{61.27}$$

Using the assumption (61.25), $\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta} \geq 0$, and $\zeta \geq 1$, we infer that

$$a_2(v_h,\zeta v_h) = (\mu v_h + \boldsymbol{\beta}\cdot\nabla v_h, \zeta v_h)_{L^2(D)} + n_{\beta h}(v_h,\zeta v_h) \geq \mu_0\|v_h\|_{L^2(D)}^2 + A_4^2,$$

since integration by parts shows that $(\mu v_h + \boldsymbol{\beta}\cdot\nabla v_h, \zeta v_h)_{L^2(D)} = ((\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta})\zeta v_h, v_h)_{L^2(D)} - \frac{1}{2}((\boldsymbol{\beta}\cdot\nabla\zeta)v_h, v_h)_{L^2(D)} + (\frac{1}{2}(\boldsymbol{\beta}\cdot\boldsymbol{n})\zeta v_h, v_h)_{L^2(\partial D)}$. Moreover, one can show (see Exercise 61.5) that there is $c_2 > 0$ such that

$$a_2(v_h,\varphi_h - \zeta v_h) \geq -c_2(A_1 + A_3 + A_4)\times\mu_0^{\frac{1}{2}}\|v_h\|_{L^2(D)}. \tag{61.28}$$

Adding the above four inequalities shows that

$$\begin{aligned} a_{\epsilon h}(v_h,\varphi_h) \geq{}& \frac{1}{2}(A_1^2 + A_2^2) + A_3^2 + A_4^2 + \mu_0\|v_h\|_{L^2(D)}^2 \\ &- (c_1\varpi_0^{\frac{1}{2}}(A_1 + A_2 + A_3) + c_2(A_1 + A_3 + A_4))\mu_0^{\frac{1}{2}}\|v_h\|_{L^2(D)}. \end{aligned}$$

Invoking Young's inequality implies that

$$\begin{aligned} a_{\epsilon h}(v_h,\varphi_h) &\geq \frac{1}{2}(A_1^2+A_2^2)+A_3^2+A_4^2+\frac{1}{2}\mu_0\|v_h\|_{L^2(D)}^2 \\ &\quad - 3(c_1^2\varpi_0+c_2^2)(A_1^2+A_3^2)-3c_1^2\varpi_0A_2^2-3c_2^2A_4^2. \end{aligned}$$

Finally, we set $w_h := \lambda v_h + \varphi_h \in V_h$ with $\lambda := 6(c_1^2\varpi_0 + c_2^2)$. This yields

$$\begin{aligned} 2a_{\epsilon h}(v_h,w_h) &= 2\lambda a_{\epsilon h}(v_h,v_h)+2a_{\epsilon h}(v_h,\varphi_h) \\ &\geq A_1^2+A_2^2+2A_3^2+2A_4^2+\mu_0\|v_h\|_{L^2(D)}^2 \\ &\quad + (\lambda-6(c_1^2\varpi_0+c_2^2))A_1^2+(\lambda-6c_1^2\varpi_0)A_2^2 \\ &\quad + (2\lambda-6(c_1^2\varpi_0+c_2^2))A_3^2+(2\lambda-6c_2^2)A_4^2 \\ &\geq A_1^2+A_2^2+2A_3^2+2A_4^2+\mu_0\|v_h\|_{L^2(D)}^2. \end{aligned}$$

Since $\|v_h\|_{V_h}^2 = A_1^2+A_2^2+A_3^2+A_4^2+\mu_0\|v_h\|_{L^2(D)}^2$, this proves that $a_{\epsilon h}(v_h,w_h) \geq \frac{1}{2}\|v_h\|_{V_h}^2$, and the conclusion follows from the bound $\|w_h\|_{V_h} \leq c\|v_h\|_{V_h}$. □

Remark 61.16 (Literature). We refer the reader to Devinatz et al. [104], Azerad [18], Ayuso and Marini [17], Deuring et al. [103], Cantin [79], Cantin and Ern [80], Bensalah et al. [30] for further results on divergence-free advection. □

Exercises

Exercise 61.1 (A priori estimates). Consider the problem (61.1). Assume that $\mathbb{d}_\epsilon := \epsilon\mathbb{I}_d$, $\nabla\cdot\boldsymbol{\beta} = 0$, $\boldsymbol{\beta}_{|\partial D} = \mathbf{0}$, $\mu := \mu_0 \geq 0$, and $f \in H_0^1(D)$. Let $\nabla_{\mathrm{s}}\boldsymbol{\beta} := \frac{1}{2}(\nabla\boldsymbol{\beta} + (\nabla\boldsymbol{\beta})^{\mathsf{T}})$ denote the symmetric part of the gradient of $\boldsymbol{\beta}$, and assume that there is $\mu_0' > 0$ s.t. $\nabla_{\mathrm{s}}\boldsymbol{\beta} + \mu\mathbb{I}_d \geq \mu_0'\mathbb{I}_d$ in the sense of quadratic forms. Prove that $|u|_{H^1(D)} \leq (\mu_0'+\mu_0)^{-1}|f|_{H^1(D)}$ and $\|\Delta u\|_{L^2(D)} \leq (4(\mu_0'+\mu_0)\epsilon)^{-\frac{1}{2}}|f|_{H^1(D)}$. (*Hint*: test the PDE (61.1) with $-\Delta u$.) *Note*: see also Beirão da Veiga [29], Burman [60].

Exercise 61.2 (Advection-diffusion, 1D). Let $D := (0,1)$ and let ϵ, b be two positive real numbers. Let $f : D \to \mathbb{R}$ be a smooth function. Consider the PDE $-\epsilon u'' + bu' = f$ in D with the boundary conditions $u(0) = 0$, $u(1) = 0$. Consider H^1-conforming $\mathbb{P}_1$ Lagrange finite elements on the uniform grid $\mathcal{T}_h$ with nodes $x_i := ih$, $\forall i \in \{0{:}I\}$, and meshsize $h := \frac{1}{I+1}$. (i) Evaluate the stiffness matrix. (*Hint*: factor out the ratio $\frac{\epsilon}{h}$ and introduce the local Péclet number $\gamma := \frac{bh}{\epsilon}$.) (ii) Solve the linear system when $f := 1$ and plot the solutions for $h := 10^{-2}$ and $\gamma \in \{0.1, 1, 10\}$. (*Hint*: the solution $\mathsf{U} \in \mathbb{R}^I$ has the form $\mathsf{U}^0 + \tilde{\mathsf{U}}$ with $\mathsf{U}_i^0 := b^{-1}ih$ and $\tilde{\mathsf{U}}_i := \varrho + \theta\delta^i$ for some constants ϱ, θ, δ.) (iii) Consider now the boundary conditions $u(0) = 0$ and $u'(1) = 0$. Write the weak formulation and show well-posedness. Evaluate the stiffness matrix. (*Hint*: this matrix is now of order $(I+1)$.) Derive the equation satisfied by $h^{-1}(\mathsf{U}_{I+1} - \mathsf{U}_I)$, and comment on the limit values obtained as $h \to 0$ with fixed $\epsilon > 0$ and as $\epsilon \to 0$ with fixed $h \in \mathcal{H}$.

Exercise 61.3 (Artificial viscosity). Consider the model problem (61.1) with $\mathbb{d} := \epsilon\mathbb{I}_d$ with constant $\epsilon > 0$. Assume that $u \in H^2(D)$. Assume that $\boldsymbol{\beta}$ is divergence-free and $\mu > 0$ is constant, and set $b := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$. Consider the finite element space $V_h := P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ on a mesh from a quasi-uniform sequence (for simplicity). Consider the following nonconsistent approximation: Find

$u_h \in V_h$ such that $a_\epsilon(u_h, w_h) + s_h(u_h, w_h) = (f, w_h)_{L^2(D)}$ for all $w_h \in V_h$, where $s_h(v_h, w_h) := \frac{1}{2}bh(\nabla v_h, \nabla w_h)_{\boldsymbol{L}^2(D)}$ for all $v_h, w_h \in P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$. (i) Prove the following error estimate:

$$\mu^{\frac{1}{2}}\|u - u_h\|_{L^2(D)} + (\epsilon^{\frac{1}{2}} + (bh)^{\frac{1}{2}})\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \le c\,(\epsilon^{\frac{1}{2}} + (bh)^{\frac{1}{2}} + \mu^{\frac{1}{2}}h + \mu^{-\frac{1}{2}}b)h|u|_{H^2(D)}.$$

(*Hint*: use the norms $\|v\|^2_{V_\flat} := (\epsilon+\frac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)}+\mu\|v\|^2_{L^2(D)}$, $\|v\|^2_{V_\sharp} := (\epsilon+\frac{1}{2}bh)\|\nabla v\|^2_{\boldsymbol{L}^2(D)}+(\mu+2bh^{-1})\|v\|^2_{L^2(D)}$ and adapt the proof of Lemma 27.8.) (ii) Consider the 1D setting of Exercise 61.2 with $f := 1$. Set $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h) = \mathrm{span}\{\varphi_i\}_{i\in\{1:I\}}$, where the φ_i's are the usual hat basis functions in $P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$. Let $\xi : [0,1] \to \mathbb{R}$ be a smooth function, called bubble function, s.t. $\xi(0) = \xi(1) = 0$ and $\xi \ge 0$. For all $i \in \{1{:}I\}$, set $\xi_i(x) := \xi(\frac{x-x_{i-1}}{h})$ if $x \in [x_{i-1}, x_i]$, $\xi_i(x) := -\xi(\frac{x-x_i}{h})$ if $x \in [x_i, x_{i+1}]$, and $\xi_i(x) := 0$ otherwise, and set $\psi_i := \varphi_i + \xi_i$. Let $W_h = \mathrm{span}\{\psi_i\}_{i\in\{1:I\}}$. Prove that the Petrov–Galerkin formulation using the pair (V_h, W_h) as trial and test spaces is equivalent to a Galerkin formulation in V_h with the bilinear form augmented by an artificial viscosity term. (*Hint*: verify that $\int_{x_{i-1}}^{x_{i+1}} u_h'\xi_i\,\mathrm{d}x = h(\int_0^1 \xi(x)\,\mathrm{d}x)\int_{x_{i-1}}^{x_{i+1}} u_h'\varphi_i'\,\mathrm{d}x$ for all $i \in \{1{:}I\}$.) Explain how to choose $\int_0^1 \xi(x)\,\mathrm{d}x$ so that the stiffness matrix is always an M-matrix. (*Hint*: use Exercise 61.2.)

Exercise 61.4 (Bound on consistency term). Prove Lemma 61.7. (*Hint*: observe that $|\boldsymbol{n}\cdot\mathrm{d}_\epsilon\nabla v_h| \le \lambda_F^{\frac{1}{2}}\|\mathrm{d}_\epsilon^{\frac{1}{2}}\nabla v_h\|_{\ell^2(\mathbb{R}^d)}$, use that $\mathrm{d}_\epsilon^{\frac{1}{2}}\nabla v_h$ is a piecewise polynomial, and adapt the proof of Lemma 37.2.)

Exercise 61.5 (Divergence-free advection). (i) Prove (61.27). (*Hint*: use Lemma 22.3 and $[\![\zeta_0 v_h]\!] = [\![\zeta_0]\!]v_h$, and bound $[\![\zeta_0]\!]$ using L_ζ.) (ii) Prove (61.28). (*Hint*: use that $\|\varphi_h - \zeta v_h\|_{L^2(K)} \le \|\varphi_h - \zeta v_h\|_{L^2(K)} + \|(\zeta - \zeta_0)v_h\|_{L^2(K)}$.) (iii) Prove that $\|\varphi_h\|_{V_h} \le c\|v_h\|_{V_h}$. (*Hint*: bound $\|\zeta_0 v_h\|_{V_h}$ and $\|\varphi_h - \zeta_0 v_h\|_{V_h}$.)

Chapter 62

Stokes equations: Residual-based stabilization

Employing inf-sup stable mixed finite elements to solve Stokes-like problems may seem to be a cumbersome constraint. The goal of this chapter is to show that it is possible to work with pairs of finite elements that do not satisfy the inf-sup condition (53.15) provided the Galerkin formulation is slightly modified. This is done by extending the stabilization techniques that have been presented in Chapters 57–60 to the Stokes problem. Although all these techniques can be adapted to the Stokes problem, for brevity we only exemplify three of them. We focus on the Galerkin/least-squares (GaLS) in this chapter. The continuous interior penalty and the discontinuous Galerkin methods are investigated in Chapter 63. The reader is referred to Braack et al. [42] for a review of stabilization techniques for the Stokes equations.

62.1 Model problem

Let D be a Lipschitz polyhedron in $\mathbb{R}^d$. As in Chapter 53, we consider the Stokes problem with homogeneous mixed Dirichlet/Neumann boundary conditions:

$$\nabla\cdot\mathbb{r}(\boldsymbol{u},p) = \boldsymbol{f}, \quad \nabla\cdot\boldsymbol{u} = g \qquad \text{in } D, \tag{62.1a}$$

$$\boldsymbol{u}_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}, \quad \mathbb{r}(\boldsymbol{u},p)_{|\partial D_{\mathrm{n}}}\boldsymbol{n} = \boldsymbol{0} \quad \text{on } \partial D, \tag{62.1b}$$

where the body force $\boldsymbol{f}$ and the mass production rate g are assumed to be in $\boldsymbol{L}^2(D)$ and $L^2(D)$, respectively. Moreover, $\mathbb{r}(\boldsymbol{u},p) := -\mathbb{s}(\boldsymbol{u}) + p\mathbb{I}$ is the total stress tensor, $\mathbb{s}(\boldsymbol{u}) := 2\mu\mathbb{e}(\boldsymbol{u})$ the viscous stress tensor, $\mathbb{e}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathsf{T}})$ the (linearized) strain rate tensor, and $\mu > 0$ the dynamic viscosity. For simplicity, we assume that μ is constant and that $|\partial D_{\mathrm{d}}| > 0$. We consider the functional space $Y := \boldsymbol{V}_{\mathrm{d}}\times Q$ with

$$\boldsymbol{V}_{\mathrm{d}} := \{\boldsymbol{v}\in\boldsymbol{H}^1(D) \mid \gamma^{\mathrm{g}}(\boldsymbol{v})_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}\}, \tag{62.2}$$

$$Q := \begin{cases} L^2(D) & \text{if } \partial D \neq \partial D_{\mathrm{d}}, \\ L^2_*(D) := \{q\in L^2(D) \mid \int_D q\,\mathrm{d}x = 0\} & \text{if } \partial D = \partial D_{\mathrm{d}}, \end{cases} \tag{62.3}$$

with the trace map $\gamma^{\mathrm{g}} : \boldsymbol{H}^1(D) \to \boldsymbol{H}^{\frac{1}{2}}(\partial D)$. We define the bilinear forms $a : \boldsymbol{V}_{\mathrm{d}}\times\boldsymbol{V}_{\mathrm{d}} \to \mathbb{R}$, $b : \boldsymbol{V}_{\mathrm{d}}\times Q \to \mathbb{R}$ s.t. $a(\boldsymbol{v},\boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}){:}\mathbb{e}(\boldsymbol{w})\,\mathrm{d}x$, $b(\boldsymbol{v},q) := -\int_D q\nabla\cdot\boldsymbol{v}\,\mathrm{d}x$, and combine them into

the bilinear form $t : Y \times Y \to \mathbb{R}$ s.t.

$$t((\boldsymbol{v},q),(\boldsymbol{w},r)) := a(\boldsymbol{v},\boldsymbol{w}) + b(\boldsymbol{w},q) - b(\boldsymbol{v},r). \tag{62.4}$$

Setting $\ell(\boldsymbol{w},r) := \int_D \boldsymbol{f}\cdot\boldsymbol{w}\,\mathrm{d}x + \int_D gr\,\mathrm{d}x$, it has been established in Theorem 53.11 that the following problem is well-posed:

$$\begin{cases} \text{Find } (\boldsymbol{u},p) \in Y \text{ such that} \\ t((\boldsymbol{u},p),(\boldsymbol{w},r)) = \ell(\boldsymbol{w},r), \quad \forall (\boldsymbol{w},r) \in Y. \end{cases} \tag{62.5}$$

In particular, Lemma 53.12 shows that the following inf-sup condition on the bilinear form t holds true uniformly w.r.t. $\mu > 0$:

$$\inf_{(\boldsymbol{v},q)\in Y} \sup_{(\boldsymbol{w},r)\in Y} \frac{|t((\boldsymbol{v},q),(\boldsymbol{w},r))|}{\|(\boldsymbol{v},q)\|_Y \|(\boldsymbol{w},r)\|_Y} =: \gamma > 0, \tag{62.6}$$

with the norm $\|(\boldsymbol{v},q)\|_Y^2 := \mu|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2 + \mu^{-1}\|q\|_{L^2(D)}^2$.

We want to construct approximation methods for the solution to (62.5) with discrete velocity spaces $\boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_{\mathrm{d}}$ and discrete pressure spaces $Q_h \subset Q$ that do not satisfy the inf-sup condition (53.15). Letting $Y_h := \boldsymbol{V}_{h\mathrm{d}} \times Q_h$, the central idea of this chapter and the next one is to modify the bilinear form t by adding some stabilization terms so as to produce a discrete bilinear form t_h satisfying an inf-sup condition on $Y_h \times Y_h$ uniformly w.r.t. $h \in \mathcal{H}$ and $\mu > 0$.

62.2 Discrete setting for GaLS stabilization

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular mesh sequence s.t. for all $h \in \mathcal{H}$, $\mathcal{T}_h$ covers D exactly and each boundary face of $\mathcal{T}_h$ is either in ∂D_{d} or in ∂D_{n}. The sets of the boundary faces in ∂D_{d} and ∂D_{n} are denoted by $\mathcal{F}_h^{\mathrm{d}}$ and $\mathcal{F}_h^{\mathrm{n}}$, respectively. Let $\boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_{\mathrm{d}}$ and $Q_h \subset Q$ be two finite element spaces constructed on the mesh $\mathcal{T}_h$. Notice that the approximation setting is conforming. In particular, the velocity approximation is $\boldsymbol{H}^1$-conforming, and the Dirichlet condition is strongly enforced. The discrete pressures can be continuous or discontinuous. The examples we have in mind are

$$\boldsymbol{V}_{h\mathrm{d}} := \boldsymbol{P}_{k_u}^{\mathrm{g}}(\mathcal{T}_h) \cap \boldsymbol{V}_{\mathrm{d}}, \qquad Q_h := P_{k_p}^{\mathrm{b}}(\mathcal{T}_h) \text{ or } Q_h := P_{k_p}^{\mathrm{g}}(\mathcal{T}_h), \tag{62.7}$$

with $k_u \geq 1$, and either $k_p \geq 0$ if $Q_h := P_{k_p}^{\mathrm{b}}(\mathcal{T}_h)$ or $k_p \geq 1$ if $Q_h := P_{k_p}^{\mathrm{g}}(\mathcal{T}_h)$.

Remark 62.1 (Pressure space). If $\partial D = \partial D_{\mathrm{d}}$, it is implicitly understood that the discrete pressure space incorporates the zero mean-value condition. To simplify the notation, this condition is not stated explicitly. In practice, the discrete problem can be assembled without enforcing this condition since it can be easily handled when solving the linear system. □

Recall that $Y_h := \boldsymbol{V}_{h\mathrm{d}} \times Q_h$. We construct a GaLS approximation of (62.5) by proceeding similarly to §57.3. The general idea is to add suitable least-squares (LS) penalties to the bilinear form t in order to obtain a discrete bilinear form $t_h : Y_h \times Y_h \to \mathbb{R}$ satisfying the following inf-sup condition:

$$\inf_{(\boldsymbol{v}_h,q_h)\in Y_h} \sup_{(\boldsymbol{w}_h,r_h)\in Y_h} \frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{v}_h,q_h)\|_{Y_h} \|(\boldsymbol{w}_h,r_h)\|_{Y_h}} \geq \gamma_0 > 0, \tag{62.8}$$

where $\|\cdot\|_{Y_h}$ is a discrete counterpart of $\|\cdot\|_Y$ and γ_0 is bounded away from 0 for all $h \in \mathcal{H}$ and all $\mu > 0$. The above goal is reached by introducing the stabilized bilinear form $t_h : Y_h \times Y_h \to \mathbb{R}$ defined by

$$t_h(x_h,y_h) := t(x_h,y_h) + s_h(x_h,y_h), \qquad s_h := s_h^{\mathrm{r}} + s_h^{\mathrm{p}} + s_h^{\mathrm{n}}, \tag{62.9}$$

with $s_h^{\mathrm{r}}, s_h^{\mathrm{p}}$, and s_h^{n} s.t. for all $x_h, y_h \in Y_h$ and all $q_h, r_h \in Q_h$,

$$s_h^{\mathrm{r}}(x_h, y_h) \coloneqq \sum_{K\in\mathcal{T}_h} \varpi^{\mathrm{r}} \frac{h_K^2}{\mu} (\nabla_h{\cdot}\mathbb{r}(x_h), \nabla_h{\cdot}\mathbb{r}(y_h))_{\boldsymbol{L}^2(K)}, \tag{62.10a}$$

$$s_h^{\mathrm{p}}(q_h, r_h) \coloneqq \sum_{F\in\mathcal{F}_h^\circ} \varpi^{\mathrm{p}} \frac{h_F}{\mu} (\llbracket q_h \rrbracket, \llbracket r_h \rrbracket)_{L^2(F)}, \tag{62.10b}$$

$$s_h^{\mathrm{n}}(x_y, y_h) \coloneqq \sum_{F\in\mathcal{F}_h^{\mathrm{n}}} \varpi^{\mathrm{n}} \frac{h_F}{\mu} (\mathbb{r}(x_h)\boldsymbol{n}, \mathbb{r}(y_h)\boldsymbol{n})_{\boldsymbol{L}^2(F)}, \tag{62.10c}$$

where $\varpi^{\mathrm{r}}, \varpi^{\mathrm{p}}, \varpi^{\mathrm{n}}$ are nondimensional constants of order 1, and $\nabla_h{\cdot}$ denotes the broken divergence operator, i.e., $(\nabla_h{\cdot}\mathbb{r})_{|K} \coloneqq \nabla{\cdot}(\mathbb{r}_{|K})$ for all $K \in \mathcal{T}_h$ (recall that $\mathbb{r}(\boldsymbol{v}_h, q_h)$ does not have a weak divergence in $\boldsymbol{L}^2(D)$ since the normal component of $\nabla\boldsymbol{v}_h$ and the pressure q_h can jump across the mesh interfaces). After a proper modification of the right-hand side, s_h^{r} will contribute to the LS penalty on the residual $\nabla_h{\cdot}\mathbb{r}(\boldsymbol{u}_h, p_h) - \boldsymbol{f}$. Moreover, s_h^{p} is a LS penalty on the pressure jumps across the mesh interfaces and s_h^{n} is a LS penalty on the normal stress at the Neumann boundary. Obviously, s_h^{p} vanishes identically if one uses continuous discrete pressures, and s_h^{n} vanishes identically in the case of pure Dirichlet conditions, i.e., if $\partial D_{\mathrm{n}} = \emptyset$. The GaLS approximation of (62.5) is as follows:

$$\begin{cases} \text{Find } (\boldsymbol{u}_h, p_h) \in Y_h \text{ such that} \\ t_h((\boldsymbol{u}_h, p_h), (\boldsymbol{w}_h, r_h)) = \ell_h(\boldsymbol{w}_h, r_h), \quad \forall (\boldsymbol{w}_h, r_h) \in Y_h, \end{cases} \tag{62.11}$$

with

$$\ell_h(\boldsymbol{w}_h, r_h) \coloneqq \ell(\boldsymbol{w}_h, r_h) + \sum_{K\in\mathcal{T}_h} \varpi^{\mathrm{r}} \frac{h_K^2}{\mu} (\boldsymbol{f}, \nabla_h{\cdot}\mathbb{r}(\boldsymbol{w}_h, r_h))_{\boldsymbol{L}^2(K)}. \tag{62.12}$$

The last term in (62.12) ensures consistency (i.e., the Galerkin orthogonality property as shown in Theorem 62.5). To sum up, the discrete problem contains LS penalty terms on the momentum residual, on the pressure jumps across the interfaces, and on the normal force at the Neumann boundary faces.

Remark 62.2 (Literature). The idea of penalizing the residual is proposed in Hughes and Franca [189]. Other (equivalent) residual-based stabilization techniques can be constructed by playing with the structure of the stabilizing bilinear form s_h. We refer the reader to Franca and Frey [129], Tobiska and Verfürth [276], Braack et al. [42], and the references therein. In the lowest-order case with $k_u = k_p \coloneqq 1$, s_h^{r} penalizes the pressure gradient. This form of stabilization takes its origin in the work of Brezzi and Pitkäranta [53]. □

62.3 Stability and well-posedness

In order to establish the well-posedness of the discrete problem (62.11), we introduce the following norm on Y_h:

$$\|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2 \coloneqq \|(\boldsymbol{v}_h, q_h)\|_Y^2 + |(\boldsymbol{v}_h, q_h)|_S^2, \tag{62.13}$$

with $|(\boldsymbol{v}_h, q_h)|_S^2 \coloneqq s_h((\boldsymbol{v}_h, q_h), (\boldsymbol{v}_h, q_h))$. Recall that the product space $Y \coloneqq \boldsymbol{V}_{\mathrm{d}}{\times}Q$ is equipped with the norm $\|(\boldsymbol{v}, q)\|_Y^2 \coloneqq \mu|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2 + \mu^{-1}\|p\|_{L^2(D)}^2$. We also consider with obvious notation the seminorms $|\cdot|_{S^{\mathrm{r}}}$, $|\cdot|_{S^{\mathrm{p}}}$, $|\cdot|_{S^{\mathrm{n}}}$, so that $|\cdot|_S^2 = |\cdot|_{S^{\mathrm{r}}}^2 + |\cdot|_{S^{\mathrm{p}}}^2 + |\cdot|_{S^{\mathrm{n}}}^2$.

Lemma 62.3 (Stability, well-posedness). *Let t_h be defined in* (62.9) *with the stabilizing bilinear forms defined in* (62.10). (i) *There is $\gamma_0 > 0$ such that t_h satisfies the inf-sup condition*

$$\inf_{(\boldsymbol{v}_h,q_h)\in Y_h}\sup_{(\boldsymbol{w}_h,r_h)\in Y_h}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{v}_h,q_h)\|_{Y_h}\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}\geq\gamma_0>0,\tag{62.14}$$

for all $h\in\mathcal{H}$ and all $\mu>0$. (ii) *The discrete problem* (62.11) *is well-posed.*

Proof. Let $(\boldsymbol{v}_h,q_h)\in Y_h$ and $\mathbb{S}:=\sup_{(\boldsymbol{w}_h,r_h)\in Y_h}\frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}$. The proof is similar to that of the continuous inf-sup condition (62.6) (see Lemma 53.12). Since we have established in Theorem 42.10 (Korn's second inequality, see (42.14)) that $a(\boldsymbol{v},\boldsymbol{v})\geq 2\mu C_{\mathrm{K}}^2|\boldsymbol{v}|^2_{\boldsymbol{H}^1(D)}$ for all $\boldsymbol{v}\in\boldsymbol{V}_{\mathrm{d}}$, and since $\boldsymbol{V}_{h\mathrm{d}}\subset\boldsymbol{V}_{\mathrm{d}}$, letting $\alpha:=\min(1,2C_{\mathrm{K}}^2)>0$, we have

$$\alpha\big(\mu|\boldsymbol{v}_h|^2_{\boldsymbol{H}^1(D)}+|(\boldsymbol{v}_h,q_h)|_S^2\big)\leq t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{v}_h,q_h))\leq\mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}.\tag{62.15}$$

It remains to estimate $\mu^{-1}\|q_h\|^2_{L^2(D)}$. The surjectivity of the divergence operator (see Lemma 53.9) and the converse statement in Lemma C.42 imply that there exists a function $\boldsymbol{w}_{q_h}\in\boldsymbol{V}_{\mathrm{d}}$ s.t.

$$\nabla{\cdot}\boldsymbol{w}_{q_h}=-\mu^{-1}q_h\qquad\text{and}\qquad\beta_D|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\leq\mu^{-1}\|q_h\|_{L^2(D)}.\tag{62.16}$$

Let us set $\boldsymbol{w}_h:=\boldsymbol{\mathcal{I}}^{\mathrm{u}}_{h\mathrm{d}}(\boldsymbol{w}_{q_h})$, where $\boldsymbol{\mathcal{I}}^{\mathrm{u}}_{h\mathrm{d}}:\boldsymbol{V}_{\mathrm{d}}\to\boldsymbol{V}_{h\mathrm{d}}$ is the $\mathbb{R}^d$-valued version of the H^1-conforming quasi-interpolation operator $\mathcal{I}^{\mathrm{g,av}}_{h0}$ from §22.4 modified so as to satisfy the zero trace prescription on ∂D_{d}. Owing to the stability and approximation properties of $\mathcal{I}^{\mathrm{g,av}}_h$, we infer that there is c s.t. for all $h\in\mathcal{H}$, all $K\in\mathcal{T}_h$ and all $\boldsymbol{w}\in\boldsymbol{V}_{\mathrm{d}}$,

$$\|\boldsymbol{w}-\boldsymbol{\mathcal{I}}^{\mathrm{u}}_{h\mathrm{d}}(\boldsymbol{w})\|_{\boldsymbol{L}^2(K)}+h_K|\boldsymbol{w}-\boldsymbol{\mathcal{I}}^{\mathrm{u}}_{h\mathrm{d}}(\boldsymbol{w})|_{H^1(K)}\leq c\,h_K|\boldsymbol{w}|_{\boldsymbol{H}^1(D_K)},\tag{62.17}$$

where D_K is the set of the points composing the mesh cells sharing at least a vertex with K. Using the definition of the norm $\|{\cdot}\|_{Y_h}$, an inverse inequality and a discrete trace inequality to bound $|(\boldsymbol{w}_h,0)|_S$, we infer that

$$\|(\boldsymbol{w}_h,0)\|_{Y_h}\leq\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}+|(\boldsymbol{w}_h,0)|_S\leq c\,\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}.$$

We can bound $|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}$ owing to the H^1-stability of $\boldsymbol{\mathcal{I}}^{\mathrm{u}}_{h\mathrm{d}}$ from (62.17) and the regularity of the mesh sequence. Using the bound on $\boldsymbol{w}_{q_h}$ from (62.16), this yields

$$\|(\boldsymbol{w}_h,0)\|_{Y_h}\leq c\,\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\leq c'\,\mu^{-\frac12}\|q_h\|_{L^2(D)}.\tag{62.18}$$

A straightforward calculation using $(\boldsymbol{w}_h,0)$ as a test function shows that

$$\begin{aligned}\mu^{-1}\|q_h\|^2_{L^2(D)}=\;&t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))\\&-\big((\mathbb{s}(\boldsymbol{v}_h),\mathbb{e}(\boldsymbol{w}_h))_{\mathbb{L}^2(D)}+(q_h,\nabla{\cdot}(\boldsymbol{w}_{q_h}-\boldsymbol{w}_h))_{L^2(D)}\big)\\&-\big(s^{\mathrm{r}}_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))+s^{\mathrm{n}}_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0))\big).\end{aligned}$$

The rest of the proof consists of estimating the three terms on the right-hand side, say $\mathfrak{T}_1$, $\mathfrak{T}_2$, $\mathfrak{T}_3$. Owing to (62.18), we have

$$|\mathfrak{T}_1|\leq\mathbb{S}\|(\boldsymbol{w}_h,0)\|_{Y_h}\leq c\,\mathbb{S}\mu^{-\frac12}\|q_h\|_{L^2(D)}.$$

Moreover, we have $\mathfrak{T}_2=(\mathbb{s}(\boldsymbol{v}_h),\mathbb{e}(\boldsymbol{w}_{q_h}-\boldsymbol{w}_h))_{\mathbb{L}^2(D)}-(q_h,\nabla{\cdot}(\boldsymbol{w}_{q_h}-\boldsymbol{w}_h))_{L^2(D)}-(\mathbb{s}(\boldsymbol{v}_h),\mathbb{e}(\boldsymbol{w}_{q_h}))_{\mathbb{L}^2(D)}$. Integrating by parts the first two terms, we infer that $\mathfrak{T}_2=\mathfrak{T}_{2,1}+\mathfrak{T}_{2,2}+\mathfrak{T}_{2,3}+\mathfrak{T}_{2,4}$ with

$$\begin{aligned}&\mathfrak{T}_{2,1}:=(\nabla_h{\cdot}\mathbb{r}(\boldsymbol{v}_h,q_h),\boldsymbol{w}_{q_h}-\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)},&&\mathfrak{T}_{2,2}:=(\mathbb{r}(\boldsymbol{v}_h,q_h)\boldsymbol{n},\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(\partial D_{\mathrm{n}})},\\&\mathfrak{T}_{2,3}:=\sum_{F\in\mathcal{F}_h^\circ}(\llbracket\mathbb{r}(\boldsymbol{v}_h,q_h)\rrbracket\boldsymbol{n}_F,\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)},&&\mathfrak{T}_{2,4}:=-(\mathbb{s}(\boldsymbol{v}_h),\mathbb{e}(\boldsymbol{w}_{q_h}))_{\mathbb{L}^2(D)}.\end{aligned}$$

Using the Cauchy–Schwarz inequality, the definition of s_h^{r}, the approximation property (62.17), and the bound on $|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}$ in (62.16), we infer that

$$\begin{aligned}|\mathfrak{T}_{2,1}| &\le |(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}\Big(\sum_{K\in\mathcal{T}_h}(\varpi^{\mathrm{r}})^{-1}\mu h_K^{-2}\|\boldsymbol{w}_{q_h}-\boldsymbol{w}_h\|_{\boldsymbol{L}^2(K)}^2\Big)^{\frac12}\\ &\le c\,|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}\Big(\sum_{K\in\mathcal{T}_h}\mu|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D_K)}^2\Big)^{\frac12}\\ &\le c'\,|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)} \le c''\,|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{r}}}\mu^{-\frac12}\|q_h\|_{L^2(D)}.\end{aligned}$$

Proceeding similarly by using the definition of s_h^{n} yields

$$|\mathfrak{T}_{2,2}| \le c\,|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}\mu^{-\frac12}\|q_h\|_{L^2(D)},$$

where we used that $h_K^{\frac12}\|\boldsymbol{w}-\boldsymbol{\mathcal{I}}_{hd}^{\mathrm{u}}(\boldsymbol{w})\|_{\boldsymbol{L}^2(\partial K)} \le ch_K|\boldsymbol{w}|_{\boldsymbol{H}^1(D_K)}$ (which follows from (62.17) and the multiplicative trace inequality (12.15)). Concerning $\mathfrak{T}_{2,3}$, we first observe that

$$\mathfrak{T}_{2,3} = \sum_{F\in\mathcal{F}_h^\circ} -(2\mu[\![\mathfrak{e}(\boldsymbol{v}_h)]\!]\boldsymbol{n}_F,\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)} + ([\![q_h]\!]\boldsymbol{n}_F,\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)}.$$

We bound the first term involving $[\![\mathfrak{e}(\boldsymbol{v}_h)]\!]$ by means of a discrete trace inequality and the approximation property (62.17), and we proceed as above using the definition of s_h^{p} to bound the second term involving $[\![q_h]\!]$. This leads to $|\mathfrak{T}_{2,3}| \le c\,\big(\mu^{\frac12}|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}+|q_h|_{S^{\mathrm{p}}}\big)\mu^{-\frac12}\|q_h\|_{L^2(D)}$. Invoking the Cauchy–Schwarz inequality and the bound (62.16), we finally infer that $|\mathfrak{T}_{2,4}| \le c\mu^{\frac12}|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}\mu^{-\frac12}\|q_h\|_{L^2(D)}$. In summary, we have shown that

$$|\mathfrak{T}_2| \le c\,\big(\mu^{\frac12}|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}+|(\boldsymbol{v}_h,q_h)|_S\big)\mu^{-\frac12}\|q_h\|_{L^2(D)}.$$

Concerning $\mathfrak{T}_3$, since the stabilization bilinear form s_h is symmetric positive semidefinite, we have

$$|\mathfrak{T}_3| \le |(\boldsymbol{v}_h,q_h)|_S|(\boldsymbol{w}_h,0)|_S \le c\,|(\boldsymbol{v}_h,q_h)|_S\mu^{-\frac12}\|q_h\|_{L^2(D)},$$

where we used (62.18) in the last bound. Putting everything together, we can bound $\mu^{-1}\|q_h\|_{L^2(D)}^2 = \mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3$ as follows:

$$\mu^{-1}\|q_h\|_{L^2(D)}^2 \le c\,\big(\mathbb{S}+\mu^{\frac12}|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}+|(\boldsymbol{v}_h,q_h)|_S\big)\mu^{-\frac12}\|q_h\|_{L^2(D)},$$

so that $\frac{1}{\mu}\|q_h\|_{L^2(D)}^2 \le c\big(\mathbb{S}^2+\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2+|(\boldsymbol{v}_h,q_h)|_S^2\big)$. Recalling (62.15), i.e., $\big(\mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2+|(\boldsymbol{v}_h,q_h)|_S^2\big) \le \frac{1}{\alpha}\mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}$, we obtain $\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^2 \le c(\mathbb{S}^2+\mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h})$. Invoking Young's inequality, we infer that $\|(\boldsymbol{v}_h,q_h)\|_{Y_h} \le c\,\mathbb{S}$, i.e., (62.14) holds true. Finally, the well-posedness of (62.11) is a direct consequence of the inf-sup condition (62.14) combined with Theorem 26.6. □

62.4 Error analysis

The main tool to perform the error analysis is Lemma 27.5. However, the present setting allows for a slightly simpler formulation based on the Galerkin orthogonality property. Since we are going to use this property several times, we present an abstract result regarding the following generic problem: Find $y_h \in Y_h$ s.t. $t_h(y_h,z_h) = \ell_h(z_h)$ for all $z_h \in Y_h$.

Lemma 62.4 (Error estimate with Galerkin orthogonality). *Let $Y_h \subset Y$ be equipped with a norm $\|\cdot\|_{Y_h}$, let $Y_\mathrm{S} \subset Y$, and assume that $Y_\sharp := Y_\mathrm{S} + Y_h$ is equipped with a norm $\|\cdot\|_{Y_\sharp}$ that is the natural extension of the norm $\|\cdot\|_{Y_h}$ to $Y_\sharp$. Assume the following:* (i) *Stability: the bilinear form $t_h : Y_h \times Y_h \to \mathbb{R}$ satisfies the inf-sup condition* (62.14) *with the constant γ_0 for all $h \in \mathcal{H}$;* (ii) *Consistency/boundedness: Let y solve* (62.5) *and assume that $y \in V_\mathrm{S}$. Assume also that the bilinear form $t_h := t + s_h : Y_h \times Y_h \to \mathbb{R}$ can be extended to a bounded bilinear form $t_\sharp := t + s_\sharp : Y_\sharp \times Y_h \to \mathbb{R}$ with boundedness constant $\|t_\sharp\|$, and the following Galerkin orthogonality property holds true:*

$$t_\sharp(y, z_h) = \ell_h(z_h), \qquad \forall z_h \in Y_h. \tag{62.19}$$

Then the following quasi-optimal error estimate holds true:

$$\|y - y_h\|_{Y_\sharp} \le \left(1 + \frac{\|t_\sharp\|}{\gamma_0}\right) \inf_{\zeta_h \in Y_h} \|y - \zeta_h\|_{Y_\sharp}. \tag{62.20}$$

Proof. We apply Lemma 27.5. The consistency error satisfies

$$\langle \delta_h(\zeta_h), z_h \rangle_{Y_h', Y_h} = \ell_h(z_h) - t_h(\zeta_h, z_h) = t_\sharp(y - \zeta_h, z_h),$$

for all $\zeta_h, z_h \in Y_h$, where we used the Galerkin orthogonality property and the fact that $t_\sharp$ is an extension of t_h. The boundedness of $t_\sharp$ implies that $\|\delta_h(\zeta_h)\|_{Y_h'} \le \|t_\sharp\| \|y - \zeta_h\|_{Y_\sharp}$, i.e., the consistency/boundedness property (27.4) holds true with $\omega_{\sharp h} := \|t_\sharp\|$. Since $\|\cdot\|_{Y_\sharp}$ is the natural extension of $\|\cdot\|_{Y_h}$ to $Y_\sharp$, we infer that (27.5) holds true with $c_\sharp := 1$. Thus, (62.20) is just a rewriting of (27.6). □

We now apply Lemma 62.4 to the GaLS approximation of the Stokes equations. We assume that there is $r \ge 1$ s.t. the solution to (62.5) is in

$$Y_\mathrm{S} := \{(\boldsymbol{v}, q) \in Y \mid (\boldsymbol{v}, q) \in \boldsymbol{H}^{1+r}(D) \times H^r(D),\ \nabla\cdot\mathbb{r}(\boldsymbol{v}, q) \in \boldsymbol{L}^2(D)\}, \tag{62.21}$$

and we set $Y_\sharp := Y_\mathrm{S} + Y_h$. Notice that the solution to (62.5) satisfies $\nabla\cdot\mathbb{r}(\boldsymbol{u}, p) \in \boldsymbol{L}^2(D)$ by assumption since $\boldsymbol{f} \in \boldsymbol{L}^2(D)$. Let $\|\cdot\|_{Y_\sharp}$ be the natural extension to $Y_\sharp$ of the norm $\|\cdot\|_{Y_h}$ defined in (62.13), and let $t_\sharp := t + s_\sharp$ be the natural extension to $Y_\sharp \times Y_h$ of the bilinear form $t_h := t + s_h$ defined in (62.9). With these extensions, $s_h((\boldsymbol{v}, q), (\boldsymbol{w}_h, r_h))$ is well defined for all $(\boldsymbol{w}_h, r_h) \in Y_h$ (see (62.10)), since the assumption $r \ge 1$ implies that for all $(\boldsymbol{v}, q) \in Y_\sharp$, $\nabla\cdot\mathbb{r}(\boldsymbol{v}, q) \in \boldsymbol{L}^2(D)$, and the normal trace of $\mathbb{r}(\boldsymbol{v}, q)$ and the pressure jumps are well defined.

Theorem 62.5 (Error estimate). *Let $(\boldsymbol{u}, p)$ solve* (62.5) *and assume that $(\boldsymbol{u}, p) \in Y_\mathrm{S}$ as defined in* (62.21). *Let $(\boldsymbol{u}_h, p_h) \in Y_h$ solve* (62.11) *with the stabilizing bilinear forms defined in* (62.10). (i) *There is c such that for all $h \in \mathcal{H}$ and all $\mu > 0$,*

$$\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)\|_{Y_\sharp} \le c \inf_{(\boldsymbol{v}_h, q_h) \in Y_h} \|(\boldsymbol{u} - \boldsymbol{v}_h, p - q_h)\|_{Y_\sharp}. \tag{62.22}$$

(ii) *Assuming $(\boldsymbol{u}, p) \in \boldsymbol{H}^{1+\tau}(D) \times H^\tau(D)$ with $\tau \in [1, \min(k_u, k_p + 1)]$, the following holds true:*

$$\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)\|_{Y_\sharp} \le c \left(\sum_{K \in \mathcal{T}_h} \mu h_K^{2\tau} |\boldsymbol{u}|^2_{\boldsymbol{H}^{1+\tau}(K)} + \frac{h_K^{2\tau}}{\mu} |p|^2_{H^\tau(K)} \right)^{\frac{1}{2}}. \tag{62.23}$$

Proof. (i) We apply Lemma 62.4 to prove (62.22). Stability using the norm $\|\cdot\|_{Y_h}$ has been established in Lemma 62.3, and one readily verifies that the extended bilinear form $t_\sharp$ is bounded on $Y_\sharp \times Y_h$. Thus, it only remains to verify the Galerkin orthogonality property. Since the solution

to (62.5) satisfies $t((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = \ell(\boldsymbol{w}_h,r_h)$ for all $(\boldsymbol{w}_h,r_h)\in Y_h\subset Y$, $\nabla{\cdot}\mathbb{r}(\boldsymbol{u},p)=\boldsymbol{f}$ in D, $\mathbb{r}(\boldsymbol{u},p)\boldsymbol{n}=\boldsymbol{0}$ on ∂D_{n}, and $[\![p]\!]_F=0$ for all $F\in\mathcal{F}_h^\circ$ (recall that we assumed $r\geq 1$), we infer that

$$\begin{aligned}t_\sharp((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) &= t((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) + \sum_{K\in\mathcal{T}_h}\varpi^{\mathrm{r}}\frac{h_K^2}{\mu}(\boldsymbol{f},\nabla_h{\cdot}\mathbb{r}(\boldsymbol{w}_h,r_h))_{\boldsymbol{L}^2(K)}\\ &= \ell(\boldsymbol{w}_h,r_h)+\sum_{K\in\mathcal{T}_h}\varpi^{\mathrm{r}}\frac{h_K^2}{\mu}(\boldsymbol{f},\nabla_h{\cdot}\mathbb{r}(\boldsymbol{w}_h,r_h))_{\boldsymbol{L}^2(K)} = \ell_h(\boldsymbol{w}_h,r_h).\end{aligned}$$

This completes the proof of (62.22).

(ii) We consider the quasi-interpolation operator $\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}:\boldsymbol{V}_{\mathrm{d}}\to\boldsymbol{V}_{h\mathrm{d}}$ from the proof of Lemma 62.3, together with the operator $\mathcal{I}_h^{\mathrm{p}}:Q\to Q_h$ which is either the H^1-conforming quasi-interpolation operator $\mathcal{I}_h^{\mathrm{g,av}}$ from §22.3 if one uses H^1-conforming discrete pressures, or the broken interpolation operator $\mathcal{I}_h^\sharp$ from §18.3 if one uses discontinuous pressures. One can then invoke the following approximation properties:

$$|\boldsymbol{w}-\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{w})|_{H^1(K)}+h_K|\boldsymbol{w}-\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{w})|_{H^2(K)}\leq c\,h_K|\boldsymbol{w}|_{\boldsymbol{H}^2(D_K)},\tag{62.24a}$$

$$\|q-\mathcal{I}_h^{\mathrm{p}}(q)\|_{L^2(K)}+h_K|q-\mathcal{I}_h^{\mathrm{p}}(q)|_{H^1(K)}\leq c\,h_K|q|_{H^1(D_K)},\tag{62.24b}$$

for all $K\in\mathcal{T}_h$, all $h\in\mathcal{H}$, all $\boldsymbol{w}\in\boldsymbol{H}^2(D)\cap\boldsymbol{V}_{\mathrm{d}}$, and all $q\in H^1(D)\cap M$ (D_K can be replaced by K in (62.24b) if one uses discontinuous pressures). Estimating $\|(\boldsymbol{u}-\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{u}),p-\mathcal{I}_h^{\mathrm{p}}(p))\|_Y$ is straightforward, and we refer the reader to Exercise 62.3 for the bound on the stabilization terms. □

Remark 62.6 (Optimality). The error estimate (62.23) is optimal if $\tau=k_u=k_p+1$. The simplest example is $k_u:=1$, $k_p:=0$ leading to first-order convergence rates if $(\boldsymbol{u},p)\in\boldsymbol{H}^2(D)\times H^1(D)$. □

We finish by establishing an L^2-error estimate on the velocity by using the Aubin–Nitsche duality argument. Let $s\in(0,1]$ be the regularity pickup index in the Stokes problem, i.e., there is $c>0$ s.t. for all $\boldsymbol{j}\in\boldsymbol{L}^2(D)$, the unique solution $(\boldsymbol{\xi},\zeta)\in Y$ to the problem $\nabla{\cdot}(\mathbb{r}(\boldsymbol{\xi},\zeta))=\boldsymbol{j}$, $\nabla{\cdot}\boldsymbol{\xi}=0$, and $\mathbb{r}(\boldsymbol{\xi},\zeta)_{|\partial D_{\mathrm{n}}}\boldsymbol{n}=\boldsymbol{0}$ is in $\boldsymbol{H}^{1+s}(D)\times H^s(D)$ and $\mu|\boldsymbol{\xi}|^2_{\boldsymbol{H}^{1+s}(D)}+\mu^{-1}|\zeta|^2_{H^s(D)}\leq c\mu^{-1}\ell_D^{2(1-s)}\|\boldsymbol{j}\|^2_{\boldsymbol{L}^2(D)}$, where ℓ_D is a length scale associated with D, e.g., $\ell_D:=\mathrm{diam}(D)$.

Corollary 62.7 (Velocity L^2-estimate). *Consider the setting of Theorem* 62.5. *Assume that the regularity pickup index in the Stokes problem is $s=1$. There is c such that for all $h\in\mathcal{H}$ and all $\mu>0$,*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)}\leq c\,\mu^{-\frac12}h\|(\boldsymbol{u}-\boldsymbol{u}_h,p-p_h)\|_{Y_\sharp}.$$

Proof. Let $(\boldsymbol{\eta},\delta)\in Y$ be the solution to the dual problem

$$\nabla{\cdot}(\mathbb{r}(\boldsymbol{\eta},\delta))=\boldsymbol{u}-\boldsymbol{u}_h,\quad \nabla{\cdot}\boldsymbol{\eta}=0,\quad \mathbb{r}(\boldsymbol{\eta},\delta)_{|\partial D_{\mathrm{n}}}\boldsymbol{n}=\boldsymbol{0}.$$

We observe that

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(D)} &= (\boldsymbol{u}-\boldsymbol{u}_h,\nabla{\cdot}(\mathbb{r}(\boldsymbol{\eta},\delta)))_{\boldsymbol{L}^2(D)} = (\mathbb{e}(\boldsymbol{u}-\boldsymbol{u}_h),\mathbb{s}(\boldsymbol{\eta})-\delta\mathbb{I})_{\mathbb{L}^2(D)}\\ &= a(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\eta})+b(\boldsymbol{u}-\boldsymbol{u}_h,\delta) = t((\boldsymbol{u}-\boldsymbol{u}_h,p-p_h),(\boldsymbol{\eta},-\delta)),\end{aligned}$$

where we used the symmetry of the bilinear form a and the fact that $\nabla{\cdot}\boldsymbol{\eta}=0$. Using the Galerkin orthogonality property, we infer that

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(D)} &= t((\boldsymbol{u}-\boldsymbol{u}_h,p-p_h),(\boldsymbol{\eta},-\delta))-t_h((\boldsymbol{u}-\boldsymbol{u}_h,p-p_h),(\boldsymbol{v}_h,-q_h))\\ &= t((\boldsymbol{u}-\boldsymbol{u}_h,p-p_h),(\boldsymbol{\eta}-\boldsymbol{v}_h,-\delta+q_h))-s_\sharp((\boldsymbol{u}-\boldsymbol{u}_h,p-p_h),(\boldsymbol{v}_h,-q_h)),\end{aligned}$$

for all $(\boldsymbol{v}_h, q_h) \in Y_h$. Owing to the boundedness of t on $Y \times Y$ together with the symmetry and positive semidefiniteness of $s_\sharp$, we obtain

$$\begin{aligned}\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)}^2 \le c\,\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)\|_Y \|(\boldsymbol{\eta} - \boldsymbol{v}_h, \delta - q_h)\|_Y \\ + |(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)|_S |(\boldsymbol{v}_h, q_h)|_S.\end{aligned}$$

Let $\boldsymbol{v}_h := \boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{\eta})$ and $q_h := \mathcal{I}_h^{\mathrm{p}}(\delta)$. Using the approximation properties of these operators, we infer that $|(\boldsymbol{v}_h, q_h)|_S \le ch(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^{1+s}(D)} + \mu^{-\frac12}|\delta|_{H^s(D)})$ (see Exercise 62.3). Recalling that $\|\cdot\|_{Y_\sharp} := (\|\cdot\|_Y^2 + |\cdot|_S^2)^{\frac12}$ is the natural extension of $\|\cdot\|_{Y_h}$ to $Y_\sharp$, we obtain

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)}^2 \le c\,\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)\|_{Y_\sharp} h\big(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac12}|\delta|_{H^1(D)}\big).$$

We conclude by invoking the bound

$$\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac12}|\delta|_{H^1(D)} \le c\mu^{-\frac12}\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)},$$

which follows from our assumption on the regularity pickup index. □

Exercises

Exercise 62.1 (Pressure gradient). Assume (62.14). Prove an inf-sup condition similar to (62.14) using the norm $\|(\boldsymbol{v}_h, q_h)\|_{Y_h^+}^2 := \|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2 + \sum_{K\in\mathcal{T}_h} \mu^{-1}h_K^2\|\nabla q_h\|_{\boldsymbol{L}^2(K)}^2$. (*Hint*: use an inverse inequality.)

Exercise 62.2 (Inf-sup partner). The objective of this exercise is to reprove the inf-sup condition (62.14) by identifying an inf-sup partner for all $(v_h, q_h) \in Y_h$ as suggested in Remark 25.10. (i) Prove that there is $\rho \in (0,1)$ s.t. $t_h((\boldsymbol{v}_h, q_h), ((1-\rho)\boldsymbol{v}_h + \rho\boldsymbol{w}_h, (1-\rho)q_h)) \ge \eta\|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2$ with $\boldsymbol{w}_h := \boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{w}_{q_h})$ and $\boldsymbol{w}_{q_h}$ defined in (62.16). (*Hint*: use (62.15) and the bounds on $\mathfrak{T}_2, \mathfrak{T}_3$ from the proof of Lemma 62.3.) (ii) Show that the inf-sup condition (62.14) is satisfied with a constant γ_0 depending on ρ, β_D, η, and the constant c_w introduced in (62.18), i.e., $\|(\boldsymbol{w}_h, 0)\|_{Y_h} \le c_w\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}$. (*Hint*: identify an appropriate inf-sup partner for (v_h, q_h) and use Remark 25.10.)

Exercise 62.3 (Approximation). Let $|\cdot|_S$ be the GaLS stabilization seminorm, i.e., $|\cdot|_S^2 = |\cdot|_{S^{\mathrm{r}}}^2 + |\cdot|_{S^{\mathrm{p}}}^2 + |\cdot|_{S^{\mathrm{n}}}^2$. Let $(\boldsymbol{\eta}, \zeta) \in (\boldsymbol{H}^2(D)\times H^1(D)) \cap Y$ be s.t. $\mathbb{r}(\boldsymbol{\eta}, \zeta)_{|\partial D_{\mathrm{n}}}\boldsymbol{n} = \boldsymbol{0}$. (i) Prove that $|(\boldsymbol{\eta}, \zeta)|_S \le ch(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac12}|\zeta|_{H^1(D)})$. (ii) Prove that $|(\boldsymbol{\eta} - \boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{\eta}), \zeta - \mathcal{I}_h^{\mathrm{p}}(\zeta))|_S \le ch(\mu^{\frac12}|\boldsymbol{\eta}|_{\boldsymbol{H}^2(D)} + \mu^{-\frac12}|\zeta|_{H^1(D)})$. (*Hint*: use (62.24).) (iii) Estimate $|(\boldsymbol{\mathcal{I}}_{h\mathrm{d}}^{\mathrm{u}}(\boldsymbol{\eta}), \mathcal{I}_h^{\mathrm{p}}(\zeta))|_S$.

Exercise 62.4 (Inf-sup condition on t_h). Assume that $\partial D = \partial D_{\mathrm{d}}$ so that $\boldsymbol{V}_{\mathrm{d}} := \boldsymbol{H}_0^1(D)$. Reprove (62.14) by accepting as a fact (see Exercise 63.2) that there is $\beta_0 > 0$ s.t. for all $h \in \mathcal{H}$ and all $q_h \in Q_h$,

$$\beta_0\,\mu^{-\frac12}\|q_h\|_{L^2(D)} \le \sup_{\boldsymbol{w}_h \in \boldsymbol{V}_{h\mathrm{d}}} \frac{|b(\boldsymbol{w}_h, q_h)|}{\mu^{\frac12}|\boldsymbol{w}_h|_{\boldsymbol{H}^1(D)}} + |q_h|_{S^{\mathrm{gp}}} + |q_h|_{S^{\mathrm{p}}},$$

with $|q_h|_{S^{\mathrm{gp}}}^2 := \sum_{F\in\mathcal{F}_h^\circ} \frac{h_F^3}{\mu}\|[\![\nabla_h q_h]\!]\cdot\boldsymbol{n}_F\|_{L^2(F)}^2$ for all $q_h \in Q_h$. (*Hint*: use that $b(\boldsymbol{w}_h, q_h) = t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0)) - a(\boldsymbol{v}_h, \boldsymbol{w}_h) - s_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, 0))$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}$, and prove that $|q_h|_{S^{\mathrm{gp}}}^2 \le c\big(|(\boldsymbol{v}_h, q_h)|_{S^{\mathrm{r}}}^2 + \mu|\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}^2\big)$.)

Chapter 63

Stokes equations: Other stabilizations

We continue in this chapter the study of stabilization techniques to approximate the Stokes problem (62.1) with finite element pairs that do not satisfy the inf-sup condition (53.15). We now focus our attention on the continuous interior penalty and the discontinuous Galerkin methods.

63.1 Continuous interior penalty

In this section, we take inspiration from Chapter 58 and construct a stable approximation of the Stokes problem (62.1) by replacing the control on the residual $\nabla_h \cdot \mathbb{r}(\boldsymbol{u}_h, p_h) - \boldsymbol{f}$, as done in the GaLS method studied in Chapter 62, by a control on the fluctuations of the discrete pressure. There are many ways to do that, but for the sake of brevity, we focus on a generalization of the continuous interior penalty (CIP) method presented in §58.3.

63.1.1 Discrete setting

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular mesh sequence so that each mesh covers D exactly. In order to simplify some proofs, we assume that the mesh sequence is quasi-uniform (otherwise one can use mesh-dependent weights as in §58.3). We are going to enforce weakly the Dirichlet condition on the velocity by means of Nitsche's boundary penalty method as in Chapter 37. (It is also possible to enforce strongly the velocity Dirichlet condition, but this entails distracting technicalities; see Burman and Schieweck [69]).

Let $\boldsymbol{V}_h \subset \boldsymbol{H}^1(D)$ be the discrete velocity space, i.e., the velocity approximation is H^1-conforming, but the Dirichlet condition on the velocity is not strongly enforced in $\boldsymbol{V}_h$ (i.e., $\boldsymbol{V}_h$ is not a subspace of the velocity space $\boldsymbol{V}_{\mathrm{d}}$). Although this is not a theoretical requirement, the pressure approximation is assumed to be H^1-conforming to simplify the argumentation, i.e., $Q_h \subset H^1(D)$. It is possible to consider discontinuous pressures by using the techniques presented in §63.2; see also Remark 63.5 below. The examples we have in mind are

$$\boldsymbol{V}_h := \boldsymbol{P}^{\mathrm{g}}_{k_u}(\mathcal{T}_h), \quad k_u \geq 1, \qquad Q_h := P^{\mathrm{g}}_{k_p}(\mathcal{T}_h), \quad k_p \geq 1. \tag{63.1}$$

Let us set $Y_h := \boldsymbol{V}_h \times Q_h$. If $\partial D = \partial D_{\mathrm{d}}$, it is implicitly understood that the discrete pressure space incorporates the zero mean-value condition. To simplify the notation, this condition is not stated

explicitly (see Remark 62.1). Since the Dirichlet condition on the velocity is enforced weakly, the discrete counterpart of the bilinear form a is the bilinear form $a_h : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{R}$ s.t.

$$a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) = a(\boldsymbol{v}_h, \boldsymbol{w}_h) - n_h(\boldsymbol{v}_h, \boldsymbol{w}_h) - n_h(\boldsymbol{w}_h, \boldsymbol{v}_h), \tag{63.2}$$

with $n_h(\boldsymbol{v}_h, \boldsymbol{w}_h) := \int_{\partial D_\mathrm{d}} 2\mu(\mathfrak{e}(\boldsymbol{v}_h)\boldsymbol{n})\cdot\boldsymbol{w}_h \,\mathrm{d}s$. The bilinear form b is modified accordingly, i.e., we introduce the bilinear form $b_h : \boldsymbol{V}_h \times Q_h \to \mathbb{R}$ s.t.

$$b_h(\boldsymbol{v}_h, q_h) := -\int_D (\nabla\cdot\boldsymbol{v}_h) q_h \,\mathrm{d}x + \int_{\partial D_\mathrm{d}} (\boldsymbol{v}_h\cdot\boldsymbol{n}) r_h \,\mathrm{d}s. \tag{63.3}$$

In the CIP approximation, the stabilizing bilinear form $s_h : Y_h \times Y_h \to \mathbb{R}$ is supported on $\mathcal{F}_h$ and takes the form $s_h := s_h^\mathrm{u} + s_h^\mathrm{gp} + s_h^\mathrm{n}$. The bilinear forms s_h^u and s_h^gp penalize the velocity values at the Dirichlet boundary faces and the jumps of the normal derivative of the pressure across the mesh interfaces, respectively:

$$s_h^\mathrm{u}(\boldsymbol{v}_h, \boldsymbol{w}_h) := \sum_{F\in\mathcal{F}_h^\mathrm{d}} \varpi^\mathrm{u} \frac{2\mu}{h_F} (\boldsymbol{v}_h, \boldsymbol{w}_h)_{\boldsymbol{L}^2(F)}, \tag{63.4a}$$

$$s_h^\mathrm{gp}(q_h, r_h) := \sum_{F\in\mathcal{F}_h^\circ} \varpi^\mathrm{gp} \frac{h_F^3}{\mu} (\llbracket\nabla q_h\rrbracket\cdot\boldsymbol{n}_F, \llbracket\nabla r_h\rrbracket\cdot\boldsymbol{n}_F)_{L^2(F)}, \tag{63.4b}$$

where ϖ^u, ϖ^gp are nondimensional constants, ϖ^u must be taken large enough (as usual with boundary penalty methods; see Lemma 63.2), and ϖ^gp is of order 1. The bilinear form s_h^n is meant to enforce the Neumann condition and is defined in (62.10c) ($s_h^\mathrm{n} := 0$ whenever $\partial D_\mathrm{d} = \partial D$). The CIP-stabilized bilinear form $t_h : Y_h \times Y_h \to \mathbb{R}$ is defined as

$$\begin{aligned} t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h)) := {} & a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) + b_h(\boldsymbol{w}_h, q_h) - b_h(\boldsymbol{v}_h, r_h) \\ & + s_h^\mathrm{u}(\boldsymbol{v}_h, \boldsymbol{w}_h) + s_h^\mathrm{gp}(q_h, r_h) + s_h^\mathrm{n}((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h)). \end{aligned} \tag{63.5}$$

The CIP approximation of the problem (62.5) takes the form

$$\begin{cases} \text{Find } (\boldsymbol{u}_h, p_h) \in Y_h \text{ such that} \\ t_h((\boldsymbol{u}_h, p_h), (\boldsymbol{w}_h, r_h)) = \ell_h(\boldsymbol{w}_h, r_h), \quad \forall (\boldsymbol{w}_h, r_h) \in Y_h, \end{cases} \tag{63.6}$$

with the linear form $\ell_h(\boldsymbol{w}_h, r_h) := \ell(\boldsymbol{w}_h, q_h) := \int_D \boldsymbol{f}\cdot\boldsymbol{w}_h \,\mathrm{d}x + \int_D g r_h \,\mathrm{d}x$. Notice that ℓ_h does not depend on the stabilization since the Dirichlet boundary condition on the velocity is homogeneous. To sum up, we are using LS penalties on the velocity at the Dirichlet faces, on the jumps of the normal component of the pressure gradient across the interfaces, and on the normal force at the Neumann faces.

Remark 63.1 (Literature). The present technique has been introduced in Burman and Hansbo [68], Burman et al. [72]. All the stabilization techniques presented in Chapters 58-59 can be adapted to the stabilization of the Stokes problem. Without being exhaustive, we refer the reader to Kechkar and Silvester [204], Silvester [260], Becker and Braack [28], Codina [90], Dohrmann and Bochev [108], Bochev et al. [35], Matthies et al. [226], Braack et al. [42], and the references therein. □

63.1.2 Stability and well-posedness

In order to establish the well-posedness of the discrete problem, we introduce the following norms on $\boldsymbol{V}_h$, Q_h, and Y_h:

$$\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 := 2\mu\|\mathfrak{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D)}^2 + |\boldsymbol{v}_h|_{S^{\mathrm{u}}}^2, \tag{63.7a}$$

$$\|q_h\|_{Q_h}^2 := \mu^{-1}\|q_h\|_{L^2(D)}^2 + |q_h|_{S^{\mathrm{gp}}}^2, \tag{63.7b}$$

$$\|(\boldsymbol{v}_h, q_h)\|_{Y_h}^2 := \|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 + \|q_h\|_{Q_h}^2 + |(\boldsymbol{v}_h, q_h)|_{S^{\mathrm{n}}}^2, \tag{63.7c}$$

with $|\boldsymbol{v}_h|_{S^{\mathrm{u}}}^2 := \sum_{F\in\mathcal{F}_h^{\mathrm{d}}} \frac{2\mu}{h_F}\|[\![\boldsymbol{v}_h]\!]\|_{\boldsymbol{L}^2(F)}^2$, $|(\boldsymbol{v}_h, q_h)|_{S^{\mathrm{n}}} := (s^{\mathrm{n}}((\boldsymbol{v}_h, q_h), (\boldsymbol{v}_h, q_h)))^{\frac12}$, and $|q_h|_{S^{\mathrm{gp}}}^2 := s_h^{\mathrm{gp}}(q_h, q_h)$ (we do not include the factor ϖ^{u} in $|\cdot|_{S^{\mathrm{u}}}$ to mimick the analysis of Nitsche's method for elliptic PDEs as in §37.2). Notice that $\|\cdot\|_{\boldsymbol{V}_h}$ defines a norm on $\boldsymbol{V}_h$. Indeed, if $\|\mathfrak{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D)} = |\boldsymbol{v}_h|_{S^{\mathrm{u}}} = 0$, $\boldsymbol{v}_h$ is a global rigid motion vanishing on ∂D_{d} so that $\boldsymbol{v}_h = \boldsymbol{0}$. (One can also invoke a discrete Korn's inequality; see Duarte et al. [111], Brenner [49].)

Our first step in the stability analysis is to establish coercivity for $a_h + s_h^{\mathrm{u}}$ on $\boldsymbol{V}_h$. Recall that $\mathcal{T}_h^{\partial D}$ is the collection of the mesh cells having at least one boundary face. Let n_∂ be the maximum number of boundary faces that a mesh cell in $\mathcal{T}_h^{\partial D}$ can have ($n_\partial \le d$ for simplicial meshes). Let c_{dt} be the constant so that the inverse inequality $\|\mathfrak{e}(\boldsymbol{v}_h)\boldsymbol{n}\|_{\boldsymbol{L}^2(F)} \le c_{\mathrm{dt}} h_F^{-\frac12}\|\mathfrak{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(K_l)}$ holds true for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$, all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$, and all $h \in \mathcal{H}$.

Lemma 63.2 (Coercivity of $a_h + s_h^{\mathrm{u}}$). *Assume that $\varpi^{\mathrm{u}} > n_\partial c_{\mathrm{dt}}^2$. Then setting $\alpha := \frac{\varpi^{\mathrm{u}} - n_\partial c_{\mathrm{dt}}^2}{1+\varpi^{\mathrm{u}}} > 0$, we have*

$$a_h(\boldsymbol{v}_h, \boldsymbol{v}_h) + s_h^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{v}_h) \ge \alpha\,\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2, \qquad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h. \tag{63.8}$$

Proof. See Exercise 63.1. □

Lemma 63.3 (Stability, well-posedness). *Let t_h be defined in* (63.5) *with the stabilizing bilinear forms s_h^{n} defined in* (62.10c) *and $s_h^{\mathrm{u}}, s_h^{\mathrm{gp}}$ defined in* (63.4). *Assume that $\varpi^{\mathrm{u}} > n_\partial c_{\mathrm{dt}}^2$.* (i) *There is $\gamma_0 > 0$ such that*

$$\inf_{(\boldsymbol{v}_h,q_h)\in Y_h} \sup_{(\boldsymbol{w}_h,r_h)\in Y_h} \frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{v}_h,q_h)\|_{Y_h}\|(\boldsymbol{w}_h,r_h)\|_{Y_h}} \ge \gamma_0 > 0, \tag{63.9}$$

for all $h \in \mathcal{H}$ and all $\mu > 0$. (ii) *The discrete problem* (63.6) *is well-posed.*

Proof. We only need to prove (63.9) since the well-posedness of (63.6) then follows directly. Let $(\boldsymbol{v}_h, q_h) \in Y_h$ and $\mathbb{S} := \sup_{(\boldsymbol{w}_h,r_h)\in Y_h} \frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}$. We want to show that $\gamma_0\|(\boldsymbol{v}_h,q_h)\|_{Y_h} \le \mathbb{S}$ for all $h \in \mathcal{H}$ and all $\mu > 0$. The proof is similar to that of Lemma 62.3. Since $\alpha \le 1$, Lemma 63.2 implies that

$$\alpha\big(\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 + |q_h|_{S^{\mathrm{gp}}}^2 + |(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}^2\big) \le \mathbb{S}\,\|(\boldsymbol{v}_h,q_h)\|_{Y_h}. \tag{63.10}$$

It remains to bound $\mu^{-\frac12}\|q_h\|_{L^2(D)}$. In contrast with the analysis of the GaLS stabilization, we proceed here in two steps: we first gain control on $\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)}$ and then we control $\mu^{-\frac12}\|q_h\|_{L^2(D)}$. (1) Let us first observe that the definition (63.5) of t_h implies that

$$\begin{aligned} b_h(\boldsymbol{w}_h, q_h) = {}& t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)) - a_h(\boldsymbol{v}_h,\boldsymbol{w}_h) \\ & - s_h^{\mathrm{u}}(\boldsymbol{v}_h,\boldsymbol{w}_h) - s_h^{\mathrm{n}}((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)), \end{aligned}$$

for all $\boldsymbol{w}_h \in \boldsymbol{V}_h$. Owing to the boundedness of a_h (which follows by invoking a discrete trace inequality to bound n_h), the fact that the stabilizing bilinear forms are symmetric positive semidefinite,

and the bound (63.10), we infer that

$$\begin{aligned}|b_h(\boldsymbol{w}_h,q_h)| &\le c\,\big(\mathbb{S}+\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}+|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}\big)\|(\boldsymbol{w}_h,0)\|_{Y_h}\\ &\le c'\,\big(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\big)\|(\boldsymbol{w}_h,0)\|_{Y_h}.\end{aligned}\tag{63.11}$$

Let us consider $\boldsymbol{w}_h \coloneqq \mu^{-1}h^2\mathcal{J}_h^{\mathrm{g,av}}(\nabla q_h)$, where $\mathcal{J}_h^{\mathrm{g,av}}$ is the H^1-conforming averaging operator from §22.2. Using inverse and discrete trace inequalities and the L^2-stability of $\mathcal{J}_h^{\mathrm{g,av}}$ shows that

$$\|(\boldsymbol{w}_h,0)\|_{Y_h}+\mu^{\frac12}h^{-1}\|\boldsymbol{w}_h\|_{\boldsymbol{L}^2(D)} \le c\,\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)}.$$

Since the discrete pressures are H^1-conforming, integrating by parts gives

$$\begin{aligned}\mu^{-1}h^2\|\nabla q_h\|_{\boldsymbol{L}^2(D)}^2 &= \mu^{-1}h^2(\nabla q_h-\mathcal{J}_h^{\mathrm{g,av}}(\nabla q_h),\nabla q_h)_{\boldsymbol{L}^2(D)}+b_h(\boldsymbol{w}_h,q_h)\\ &\quad+\sum_{F\in\mathcal{F}_h^{\mathrm{n}}}(q_h\boldsymbol{n},\boldsymbol{w}_h)_{\boldsymbol{L}^2(F)}.\end{aligned}$$

Notice that the contribution of the Dirichlet boundary faces is contained in the discrete bilinear form b_h. Let $\mathfrak{T}_1,\mathfrak{T}_2,\mathfrak{T}_3$ denote the three terms on the right-hand side. The crucial point is that Lemma 22.3 allows us to infer that

$$|\mathfrak{T}_1| \le c\,|q_h|_{S^{\mathrm{gp}}}\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)} \le c'\,\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)},$$

where we used (63.10) in the second bound. Using (63.11) and the above bound on $\|(\boldsymbol{w}_h,0)\|_{Y_h}$ yields

$$|\mathfrak{T}_2| \le c\,\big(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\big)\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)}.$$

Moreover, the Cauchy–Schwarz inequality and a discrete trace inequality imply that

$$|\mathfrak{T}_3| \le c\left(\sum_{F\in\mathcal{F}_h^{\mathrm{n}}}\mu^{-1}h\|q_h\boldsymbol{n}\|_{\boldsymbol{L}^2(F)}^2\right)^{\frac12}\mu^{\frac12}h^{-1}\|\boldsymbol{w}_h\|_{\boldsymbol{L}^2(D)}.$$

Since $\|q_h\boldsymbol{n}\|_{\boldsymbol{L}^2(F)} \le \|\mathbb{r}(\boldsymbol{v}_h,q_h)\boldsymbol{n}\|_{\boldsymbol{L}^2(F)}+2\mu\|\mathbb{e}(\boldsymbol{v}_h)\boldsymbol{n}\|_{\boldsymbol{L}^2(F)}$, recalling the definition of the seminorm $|\cdot|_{S^{\mathrm{n}}}$ and using a discrete trace inequality to bound $\|\mathbb{e}(\boldsymbol{v}_h)\boldsymbol{n}\|_{\boldsymbol{L}^2(F)}$, we infer that

$$\begin{aligned}|\mathfrak{T}_3| &\le c\,\big(\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}+|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}\big)\mu^{\frac12}h^{-1}\|\boldsymbol{w}_h\|_{\boldsymbol{L}^2(D)}\\ &\le c'\,\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)}.\end{aligned}$$

Putting these bounds together leads to

$$\mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)} \le c\,\big(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\big).\tag{63.12}$$

(2) Let $\boldsymbol{w}_{q_h}\in\boldsymbol{V}_{\mathrm{d}}$ be the function introduced in (62.16), i.e., $\nabla\cdot\boldsymbol{w}_{q_h}=-\mu^{-1}q_h$ and $\beta_D|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)}\le\mu^{-1}\|q_h\|_{L^2(D)}$ with $\beta_D>0$. Let us set $\boldsymbol{w}_h\coloneqq\boldsymbol{\mathcal{I}}_h^{\mathrm{g,av}}(\boldsymbol{w}_{q_h})\in\boldsymbol{V}_h$, where $\boldsymbol{\mathcal{I}}_h^{\mathrm{g,av}}$ is the $\mathbb{R}^d$-valued version of the H^1-conforming quasi-interpolation operator $\mathcal{I}_h^{\mathrm{g,av}}$ from §22.3. Notice that we have $\|(\boldsymbol{w}_h,0)\|_{Y_h}\le c\mu^{-\frac12}\|q_h\|_{L^2(D)}$. Since $\nabla\cdot\boldsymbol{w}_{q_h}=-\mu^{-1}q_h$ and the discrete pressures are H^1-conforming, integrating by parts gives

$$\begin{aligned}\mu^{-1}\|q_h\|_{L^2(D)}^2 &= (\nabla q_h,\boldsymbol{w}_{q_h}-\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}+b_h(\boldsymbol{w}_h,q_h)\\ &\quad+\sum_{F\in\mathcal{F}_h^{\mathrm{n}}}(q_h\boldsymbol{n},\boldsymbol{w}_h-\boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)} =: \mathfrak{T}_1+\mathfrak{T}_2+\mathfrak{T}_3.\end{aligned}$$

Invoking the bounds (63.12) and $\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)} \le c\mu^{-\frac12}\|q_h\|_{L^2(D)}$, the approximation properties of $\mathcal{I}_h^{\mathrm{g,av}}$, and the Cauchy–Schwarz inequality yields

$$\begin{aligned}|\mathfrak{T}_1| &\le \mu^{-\frac12}h\|\nabla q_h\|_{\boldsymbol{L}^2(D)}\mu^{\frac12}h^{-1}\|\boldsymbol{w}_{q_h}-\boldsymbol{w}_h\|_{\boldsymbol{L}^2(D)}\\ &\le c\left(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\right)\mu^{-\frac12}\|q_h\|_{L^2(D)}.\end{aligned}$$

Using (63.11) and that $\|(\boldsymbol{w}_h,0)\|_{Y_h} \le c\mu^{-\frac12}\|q_h\|_{L^2(D)}$, we obtain

$$|\mathfrak{T}_2| \le c\left(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\right)\mu^{-\frac12}\|q_h\|_{L^2(D)}.$$

Since the term $\mathfrak{T}_3$ is bounded as above, this leads to

$$\mu^{-\frac12}\|q_h\|_{L^2(D)} \le c\left(\mathbb{S}+\mathbb{S}^{\frac12}\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^{\frac12}\right).$$

We can now conclude as in the proof of Lemma 62.3. □

Remark 63.4 (Inf-sup condition on b). Assume that $\partial D = \partial D_{\mathrm{d}}$. One can show that there is $\beta_0 > 0$ s.t.

$$\beta_0\,\mu^{-\frac12}\|q_h\|_{L^2(D)} \le \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\frac{|b(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}} + |q_h|_{S^{\mathrm{GP}}}, \tag{63.13}$$

for all $q_h \in Q_h$, all $h \in \mathcal{H}$, and all $\mu > 0$; see Exercise 63.2. □

Remark 63.5 (Discontinuous pressures). Although all the arguments presented in this section are legitimate when the pressure approximation is discontinuous, say $Q_h := P_{k_p}^{\mathrm{b}}(\mathcal{T}_h)$, $k_p \ge 0$, the stabilization bilinear form s_h^{gp} is not necessary in this case. Indeed, as it will be explained for the dG formulation in §63.2, stability is then obtained by penalizing the pressure jumps across the interfaces. The lowest-order case corresponds to the $(\mathbb{P}_1,\mathbb{P}_0^{\mathrm{b}})$ pair with pressure jump stabilization. This finite element pair is investigated in Barrenechea and Valentin [24], where a divergence-free post-processing of the discrete velocity field using the lowest-order Raviart–Thomas shape functions is also proposed. □

63.1.3 Error analysis

The error analysis proceeds almost exactly as for the GaLS method. The slight difference is that now we assume that the solution to (62.5) is in

$$Y_{\mathrm{s}} := (\boldsymbol{H}^{1+r}(D)\times H^{1+r}(D))\cap Y, \qquad r > \frac12, \tag{63.14}$$

and we set as above $Y_\sharp := Y_{\mathrm{s}} + Y_h$. Compared to GaLS, the smoothness requirement on the pressure is stronger here since we need to consider the normal trace of the gradient of the pressure at the mesh interfaces. The smoothness requirement on the velocity is instead weaker since we do not need to penalize the residual in the cells. We denote by $\|\cdot\|_{Y_\sharp}$ the natural extension to $Y_\sharp$ of the norm $\|\cdot\|_{Y_h}$ defined in (63.7c) (one readily verifies that this extension indeed defines a norm). We denote by $t_\sharp$ the natural extension to $Y_\sharp\times Y_h$ of the bilinear form t_h defined in (63.5). These extensions are meaningful since $r > \frac12$.

Theorem 63.6 (Error estimate). *Let $(\boldsymbol{u},p)$ solve (62.5) and assume that $(\boldsymbol{u},p)\in Y_{\mathrm{s}}$ with Y_{s} defined in (63.14). Let $(\boldsymbol{u}_h,p_h)\in Y_h$ solve (63.6) with the stabilizing bilinear forms s_h^{n} defined*

in (62.10c) *and* $s_h^{\mathrm{u}}, s_h^{\mathrm{gp}}$ *defined in* (63.4). *Assume that* $\varpi^{\mathrm{u}} > n_\partial c_{\mathrm{dt}}^2$. (i) *There is* c *such that for all* $h \in \mathcal{H}$ *and all* $\mu > 0$,

$$\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)\|_{Y_\sharp} \le c \inf_{(\boldsymbol{v}_h,q_h)\in Y_h} \|(u-\boldsymbol{v}_h, p-q_h)\|_{Y_\sharp}. \tag{63.15}$$

(ii) *Assuming that* $k_p \ge 1$ *and recalling that* $k_u \ge 1$, *the following holds true for all* $\tau \in (\frac{1}{2}, k_u]$ *and all* $\tau' \in (\frac{3}{2}, k_p+1]$*:*

$$\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)\|_{Y_\sharp} \le c \left(\sum_{K\in\mathcal{T}_h} \mu h_K^{2\tau} |\boldsymbol{u}|^2_{\boldsymbol{H}^{1+\tau}(K)} + \frac{h_K^{2\tau'}}{\mu} |p|^2_{H^{\tau'}(K)} \right)^{\frac{1}{2}}. \tag{63.16}$$

Proof. (i) The error estimate (63.15) follows from Lemma 62.4 and Lemma 63.3 once the Galerkin orthogonality property is established. To prove the Galerkin orthogonality property (62.19), we first extend the bilinear forms to $Y_\sharp \times Y_h$ (we use a subscript $\sharp$ to denote these extensions). Let $(\boldsymbol{w}_h, r_h) \in Y_h$. We have $s_\sharp^{\mathrm{u}}(\boldsymbol{u}, \boldsymbol{w}_h) = 0$ owing to the Dirichlet boundary condition on $\boldsymbol{u}$, $s_\sharp^{\mathrm{gp}}(p, r_h) = 0$ owing to the assumed regularity on p (indeed $p \in H^{\tau'}(D)$ with $\tau' > \frac{3}{2}$ implies that $[\![\nabla p]\!]_F = \boldsymbol{0}$ for all $F \in \mathcal{F}_h^\circ$), and $s_\sharp^{\mathrm{n}}((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = 0$ owing to the Neumann boundary condition on $\mathrm{r}(\boldsymbol{u},p)$. We obtain

$$t_\sharp((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = a(\boldsymbol{u},\boldsymbol{w}_h) - n_\sharp(\boldsymbol{u},\boldsymbol{w}_h) + b_\sharp(\boldsymbol{w}_h,p) - b_\sharp(\boldsymbol{u},r_h).$$

We have

$$\begin{aligned} a_\sharp(\boldsymbol{u},\boldsymbol{w}_h) - n_\sharp(\boldsymbol{u},\boldsymbol{w}_h) &= \int_D 2\mu \mathrm{e}(\boldsymbol{u}) : \mathrm{e}(\boldsymbol{w}_h)\,\mathrm{d}x - \int_{\partial D_{\mathrm{d}}} 2\mu(\mathrm{e}(\boldsymbol{u})\boldsymbol{n})\boldsymbol{w}_h\,\mathrm{d}s,\\ b_\sharp(\boldsymbol{w}_h,p) &= -\int_D (\nabla\cdot\boldsymbol{w}_h)p\,\mathrm{d}x + \int_{\partial D_{\mathrm{d}}} (\boldsymbol{w}_h\cdot\boldsymbol{n})p\,\mathrm{d}s. \end{aligned}$$

Putting these two expressions together, integrating by parts, and using the Neumann condition satisfied by $\mathrm{r}(\boldsymbol{u},p)$, we infer that

$$a(\boldsymbol{u},\boldsymbol{w}_h) - n_\sharp(\boldsymbol{u},\boldsymbol{w}_h) + b_\sharp(\boldsymbol{w}_h,p) = (\nabla\cdot\mathrm{r}(\boldsymbol{u},p),\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\boldsymbol{f},\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}.$$

Since $b_\sharp(\boldsymbol{u},r_h) = -(g,r_h)_{L^2(D)}$, we infer that $t_\sharp((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = \ell_h(\boldsymbol{w}_h,r_h)$, i.e., the Galerkin orthogonality property holds true. Invoking Lemma 62.4 and Lemma 63.3 proves the error estimate (63.15).
(ii) The estimate (63.16) follows from (63.15) by bounding the infimum using $\boldsymbol{v}_h := \boldsymbol{\mathcal{I}}_h^{\mathrm{g,av}}(\boldsymbol{u})$, $q_h := \mathcal{I}_h^{\mathrm{g,av}}(p)$, and using the approximation properties of these quasi-interpolation operators. □

63.2 Discontinuous Galerkin

We finish our overview of stabilization methods for the Stokes problem with the discontinuous Galerkin (dG) method.

63.2.1 Discrete setting

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular mesh sequence as in §62.2. We consider the broken finite element spaces

$$\boldsymbol{V}_h := \boldsymbol{P}_{k_u}^{\mathrm{b}}(\mathcal{T}_h), \qquad Q_h := P_{k_p}^{\mathrm{b}}(\mathcal{T}_h), \qquad Y_h := \boldsymbol{V}_h \times Q_h, \tag{63.17}$$

where $k_u \geq 1$, $k_p \geq 0$, and $k_u \geq k_p$. We consider the usual average and jump operators at the mesh interfaces and extend these operators in such a way that they return the actual value of their argument at the boundary faces.

The dG counterpart of the bilinear form a is the bilinear form $a_h : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{R}$ s.t.

$$a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) := (\mathbb{s}_h(\boldsymbol{v}_h), \mathbb{e}_h(\boldsymbol{w}_h))_{\mathbb{L}^2(D)} - n_h(\boldsymbol{v}_h, \boldsymbol{w}_h) - n_h(\boldsymbol{w}_h, \boldsymbol{v}_h), \tag{63.18a}$$

$$n_h(\boldsymbol{v}_h, \boldsymbol{w}_h) := \sum_{F \in \mathcal{F}_h^{\circ} \cup \mathcal{F}_h^{\mathrm{d}}} (\{\mathbb{s}_h(\boldsymbol{v}_h)\}\boldsymbol{n}_F, [\![\boldsymbol{w}_h]\!])_{\boldsymbol{L}^2(F)}, \tag{63.18b}$$

where $\mathbb{s}_h(\boldsymbol{v}_h) := 2\mu\mathbb{e}_h(\boldsymbol{v}_h)$ and $\mathbb{e}_h(\boldsymbol{v}_h)$ denotes the broken (linearized) strain tensor s.t. $\mathbb{e}_h(\boldsymbol{v}_h) := \frac{1}{2}(\nabla_h \boldsymbol{v}_h + (\nabla_h \boldsymbol{v}_h)^{\mathsf{T}})$ and ∇_h is the broken gradient operator (see Definition 36.3). The dG counterpart of the bilinear form b is the bilinear form $b_h : \boldsymbol{V}_h \times Q_h \to \mathbb{R}$ s.t.

$$b_h(\boldsymbol{v}_h, q_h) := -(\nabla_h \cdot \boldsymbol{v}_h, q_h)_{L^2(D)} + \sum_{F \in \mathcal{F}_h^{\circ} \cup \mathcal{F}_h^{\mathrm{d}}} ([\![\boldsymbol{v}_h]\!] \cdot \boldsymbol{n}_F, \{q_h\})_{L^2(F)}, \tag{63.19}$$

where $\nabla_h \cdot$ denotes the broken divergence operator. We also introduce the stabilization bilinear form

$$s_h^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{w}_h) = \sum_{F \in \mathcal{F}_h^{\circ} \cup \mathcal{F}_h^{\mathrm{d}}} \varpi^{\mathrm{u}} \frac{2\mu}{h_F} ([\![\boldsymbol{u}_h]\!], [\![\boldsymbol{v}_h]\!])_{\boldsymbol{L}^2(F)}, \tag{63.20}$$

where ϖ^{u} is a nondimensional constant to be chosen large enough (see Lemma 63.9 below). The dG-stabilized bilinear form $t_h : Y_h \times Y_h \to \mathbb{R}$ is defined as

$$\begin{aligned} t_h((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h)) := {} & a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) + b_h(\boldsymbol{w}_h, q_h) - b_h(\boldsymbol{v}_h, r_h) \\ & + s^{\mathrm{u}}(\boldsymbol{v}_h, \boldsymbol{w}_h) + s_h^{\mathrm{p}}(q_h, r_h) + s_h^{\mathrm{n}}((\boldsymbol{v}_h, q_h), (\boldsymbol{w}_h, r_h)), \end{aligned} \tag{63.21}$$

with s_h^{u} defined in (63.20) and s_h^{p}, s_h^{n} defined in (62.10). To sum up, we introduce a LS penalty on the jumps of the velocity and the pressure across the interfaces, on the velocity values at the Dirichlet faces, and on the normal force at the Neumann faces. The dG approximation of (62.5) is

$$\begin{cases} \text{Find } (\boldsymbol{u}_h, p_h) \in Y_h \text{ such that} \\ t_h((\boldsymbol{u}_h, p_h), (\boldsymbol{w}_h, r_h)) = \ell_h(\boldsymbol{w}_h, r_h), \quad \forall (\boldsymbol{w}_h, r_h) \in Y_h, \end{cases} \tag{63.22}$$

with the linear form $\ell_h(\boldsymbol{w}_h, r_h) := \ell(\boldsymbol{w}_h, r_h) := \int_D \boldsymbol{f} \cdot \boldsymbol{w}_h \,\mathrm{d}x + \int_D g r_h \,\mathrm{d}x$.

Remark 63.7 (Numerical fluxes). Similarly to §38.4, define the lifting operator $\mathbb{L}_F^l : \boldsymbol{L}^2(F) \to P_l^{\mathrm{b}}(\mathcal{T}_h; \mathbb{R}^{d \times d})$ such that

$$(\mathbb{L}_F^l(\boldsymbol{v}), \mathbb{q}_h)_{\mathbb{L}^2(D)} = (\boldsymbol{v}, \{\mathbb{q}_h\}\boldsymbol{n}_F)_{\boldsymbol{L}^2(F)},$$

for all $\boldsymbol{v} \in \boldsymbol{L}^2(F)$, all $\mathbb{q}_h \in P_l^{\mathrm{b}}(\mathcal{T}_h; \mathbb{R}^{d \times d})$, and all $F \in \mathcal{F}_h$ with $l \in \{k_u - 1, k_u\}$ (the choice for l is discussed in Remark 38.18). Define the global lifting

$$\mathbb{L}_h^l(\boldsymbol{u}_h, p_h) := \sum_{F \in \mathcal{F}_h^{\circ} \cup \mathcal{F}_h^{\mathrm{d}}} \mathbb{L}_F^l([\![\boldsymbol{u}_h]\!]) + \sum_{F \in \mathcal{F}_h^{\mathrm{n}}} \varpi^{\mathrm{n}} \frac{h_F}{\mu} \mathbb{L}_F^l(\mathbb{r}_h(\boldsymbol{u}_h, p_h)\boldsymbol{n}).$$

Let $\mathbb{r}_h(\boldsymbol{u}_h, p_h) := -2\mu\mathbb{e}_h(\boldsymbol{u}_h) + p_h\mathbb{I}$ and define the discrete total stress tensor (compare with (38.23)):

$$\widetilde{\mathbb{r}}_h^l(\boldsymbol{u}_h, p_h) := \mathbb{r}_h(\boldsymbol{u}_h, p_h) + 2\mu\mathbb{L}_h^l(\boldsymbol{u}_h, p_h).$$

Define the momentum fluxes

$$\boldsymbol{\Phi}_F^{\mathrm{u}}(\boldsymbol{u}_h,p_h) := \begin{cases} \{\mathbb{r}_h(\boldsymbol{u}_h,p_h)\}\boldsymbol{n}_F + \varpi^{\mathrm{u}}\frac{2\mu}{h_F}[\![\boldsymbol{u}_h]\!] & \text{if } F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{d}},\\ \boldsymbol{0} & \text{if } F\in\mathcal{F}_h^{\mathrm{n}}, \end{cases}$$

and the mass fluxes

$$\Phi_F^{\mathrm{p}}(\boldsymbol{u}_h,p_h) := \begin{cases} \{\boldsymbol{u}_h\}\cdot\boldsymbol{n}_F + \varpi^{\mathrm{p}}\frac{h_F}{\mu}[\![p_h]\!] & \text{if } F\in\mathcal{F}_h^{\circ},\\ 0 & \text{if } F\in\mathcal{F}_h^{\mathrm{d}},\\ \boldsymbol{u}_h\cdot\boldsymbol{n} + \varpi^{\mathrm{n}}\frac{h_F}{\mu}\boldsymbol{n}^{\mathsf{T}}\mathbb{r}_h(\boldsymbol{u}_h,p_h)\boldsymbol{n} & \text{if } F\in\mathcal{F}_h^{\mathrm{n}}. \end{cases}$$

Let $\epsilon_{K,F} := \boldsymbol{n}_K\cdot\boldsymbol{n}_F$ for all $K\in\mathcal{T}_h$ and all $F\in\mathcal{F}_K$, where $\mathcal{F}_K$ is the collection of the mesh faces composing the boundary of K. One can verify (see Exercise 63.5) that the discrete problem (63.22) is equivalent to enforcing the following local momentum and mass balance equations: For all $\boldsymbol{\xi}\in\mathbb{P}_{k_u}$ and all $\zeta\in\mathbb{P}_{k_p}$,

$$\begin{aligned} &-(\widetilde{\mathbb{r}}_h^l(\boldsymbol{u}_h,p_h),\mathbb{e}(\boldsymbol{\xi}))_{\mathbb{L}^2(K)} + \sum_{F\in\mathcal{F}_K}\epsilon_{K,F}(\boldsymbol{\Phi}_F^{\mathrm{u}}(\boldsymbol{u}_h,p_h),\boldsymbol{\xi})_{\boldsymbol{L}^2(F)} = (\boldsymbol{f},\boldsymbol{\xi})_{\boldsymbol{L}^2(K)},\\ &-(\boldsymbol{u}_h,\nabla\zeta)_{\boldsymbol{L}^2(K)} + \sum_{F\in\mathcal{F}_K}\epsilon_{K,F}(\Phi_F^{\mathrm{p}}(\boldsymbol{u}_h,p_h),\zeta)_{L^2(F)} = (g,\zeta)_{L^2(K)}. \end{aligned}$$ □

Remark 63.8 (Literature). The material in this section is adapted from Cockburn et al. [89], Di Pietro and Ern [105, §6.1]. The bilinear form s_h^{p} penalizing the pressure jumps across the mesh interfaces can be seen as a drawback since the discrete pressure enters the discrete mass conservation equation and introduces a tighter coupling between the equations which can be cumbersome. Pressure jump stabilization can be avoided if the discrete pressure space is, loosely speaking, small enough compared to the discrete velocity space. We refer the reader to Hansbo and Larson [177], Toselli [278], Girault et al. [135], Burman and Stamm [71] for examples. □

63.2.2 Stability and well-posedness

To establish the well-posedness of the discrete problem, we introduce the following norms on $\boldsymbol{V}_h$, Q_h, and Y_h:

$$\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 := 2\mu\|\mathbb{e}_h(\boldsymbol{v}_h)\|_{\mathbb{L}^2(D)}^2 + |\boldsymbol{v}_h|_{S^{\mathrm{u}}}^2, \tag{63.23a}$$

$$\|q_h\|_{Q_h}^2 := \mu^{-1}\|q_h\|_{L^2(D)}^2 + |q_h|_{S^{\mathrm{p}}}^2, \tag{63.23b}$$

$$\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^2 := \|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 + \|q_h\|_{Q_h}^2 + |(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}^2, \tag{63.23c}$$

with $|\boldsymbol{v}_h|_{S^{\mathrm{u}}}^2 := \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{d}}}\frac{2\mu}{h_F}\|[\![\boldsymbol{v}_h]\!]\|_{\boldsymbol{L}^2(F)}^2$ and $|q_h|_{S^{\mathrm{p}}}$, $|(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}$ defined as above (we do not include the factor ϖ^{u} in $|\cdot|_{S^{\mathrm{u}}}$ to mimick the elliptic case). As for CIP, one readily verifies that $\|\cdot\|_{\boldsymbol{V}_h}$ defines a norm on $\boldsymbol{V}_h$ (one can also invoke a discrete Korn's inequality; see Duarte et al. [111], Brenner [49]).

Our first step in the stability analysis is to establish coercivity for $a_h + s_h^{\mathrm{u}}$ on $\boldsymbol{V}_h$. Let n_∂ be the maximal number of faces per mesh cell ($n_\partial \le d+1$ for simplicial meshes). Let c_{dt} be the smallest constant so that the inverse inequality $\|\mathbb{e}(\boldsymbol{v}_h)\boldsymbol{n}\|_{\boldsymbol{L}^2(F)} \le c_{\mathrm{dt}}h_F^{-\frac{1}{2}}\|\mathbb{e}(\boldsymbol{v}_h)\|_{\mathbb{L}^2(K)}$ holds true for all $\boldsymbol{v}_h\in\boldsymbol{V}_h$, all $K\in\mathcal{T}_h$, all $F\in\mathcal{F}_K$, and all $h\in\mathcal{H}$.

Lemma 63.9 (Coercivity of $a_h + s_h^{\mathrm{u}}$). *Assume that $\varpi^{\mathrm{u}} > n_\partial c_{\mathrm{dt}}^2$. Let $\alpha := \frac{\varpi^{\mathrm{u}} - n_\partial c_{\mathrm{dt}}^2}{1+\varpi^{\mathrm{u}}} > 0$. Then $a_h(\boldsymbol{v}_h,\boldsymbol{v}_h) + s_h^{\mathrm{u}}(\boldsymbol{v}_h,\boldsymbol{v}_h) \ge \alpha\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2$ for all $\boldsymbol{v}_h\in\boldsymbol{V}_h$.*

Proof. Proceed as in the proof of Lemma 38.5 and Lemma 38.6. □

Lemma 63.10 (Stability, well-posedness). *Assume that the stabilizing bilinear forms $s_h^{\mathrm{p}}, s_h^{\mathrm{n}}$ are defined in* (62.10) *and s_h^{u} is defined in* (63.20). *Assume that $\varpi^{\mathrm{u}} > n_\partial c_{\mathrm{dt}}^2$.* (i) *There is $\gamma_0 > 0$ such that*

$$\inf_{(\boldsymbol{v}_h,q_h)\in Y_h} \sup_{(\boldsymbol{w}_h,r_h)\in Y_h} \frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{v}_h,q_h)\|_{Y_h}\|(\boldsymbol{w}_h,r_h)\|_{Y_h}} \geq \gamma_0 > 0, \tag{63.24}$$

for all $h \in \mathcal{H}$ and all $\mu > 0$. (ii) *The discrete problem* (63.22) *is well-posed.*

Proof. We only need to prove (63.24) since the well-posedness of (63.22) then follows directly. Let $(\boldsymbol{v}_h, q_h) \in Y_h$ and $\mathbb{S} := \sup_{(\boldsymbol{w}_h,r_h)\in Y_h} \frac{|t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,r_h))|}{\|(\boldsymbol{w}_h,r_h)\|_{Y_h}}$. Once again the proof is similar to that of Lemma 62.3. Recalling that $\alpha \leq 1$, Lemma 63.9 implies that

$$\alpha\big(\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}^2 + |q_h|_{S^{\mathrm{p}}}^2 + |(\boldsymbol{v}_h,q_h)|_{S^{\mathrm{n}}}^2\big) \leq t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{v}_h,q_h)) \leq \mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h},$$

so that it remains to bound $\mu^{-\frac12}\|q_h\|_{L^2(D)}^2$. Let $\boldsymbol{w}_{q_h} \in \boldsymbol{V}_{\mathrm{d}}$ satisfy (62.16) and let $\boldsymbol{w}_h := \boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{w}_{q_h})$ be the L^2-orthogonal projection of $\boldsymbol{w}_{q_h}$ onto $\boldsymbol{V}_h$. Since $\|\boldsymbol{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{v})\|_{\boldsymbol{L}^2(K)} + h_K|\boldsymbol{\mathcal{I}}_h^{\mathrm{b}}(\boldsymbol{v})|_{\boldsymbol{H}^1(K)} \leq ch_K|\boldsymbol{v}|_{\boldsymbol{H}^1(K)}$ for all $K \in \mathcal{T}_h$ and all $\boldsymbol{v} \in \boldsymbol{H}^1(K)$, we infer that

$$\|(\boldsymbol{w}_h,0)\|_{Y_h} \leq c\,\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h} \leq c'\,\mu^{\frac12}|\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(D)} \leq c''\,\mu^{-\frac12}\|q_h\|_{L^2(D)}, \tag{63.25}$$

where we used a discrete trace inequality to bound $|(\boldsymbol{w}_h,0)|_{S^{\mathrm{n}}}$. Furthermore, since $\mu^{-1}\|q_h\|_{L^2(D)}^2 = -(q_h, \nabla\cdot\boldsymbol{w}_{q_h})_{L^2(D)}$, integrating by parts and since $\boldsymbol{w}_{q_h}$ vanishes on ∂D_{d}, we obtain

$$\begin{aligned}
\mu^{-1}\|q_h\|_{L^2(D)}^2 &= (\nabla_h q_h, \boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(D)} - \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{n}}} (\llbracket q_h\rrbracket \boldsymbol{n}_F, \boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)} \\
&= (\nabla_h q_h, \boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} - \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{n}}} (\llbracket q_h\rrbracket \boldsymbol{n}_F, \boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)} \\
&= b_h(\boldsymbol{w}_h,q_h) + \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{n}}} (\llbracket q_h\rrbracket \boldsymbol{n}_F, \{\boldsymbol{w}_h - \boldsymbol{w}_{q_h}\})_{\boldsymbol{L}^2(F)},
\end{aligned} \tag{63.26}$$

where $(\nabla_h q_h, \boldsymbol{w}_{q_h} - \boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = 0$ follows from $\nabla_h q_h \in \boldsymbol{V}_h$ (since $k_u \geq k_p$) and where we used the identity

$$b_h(\boldsymbol{v}_h,q_h) = (\boldsymbol{v}_h, \nabla_h q_h)_{\boldsymbol{L}^2(D)} - \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{n}}} (\{\boldsymbol{v}_h\}\cdot\boldsymbol{n}_F, \llbracket q_h\rrbracket)_{L^2(F)}, \tag{63.27}$$

for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_h \times Q_h$; see Exercise 63.4 for the proof. Using that

$$b_h(\boldsymbol{w}_h,q_h) = t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)) - a_h(\boldsymbol{v}_h,\boldsymbol{w}_h) - s_h^{\mathrm{u}}(\boldsymbol{v}_h,\boldsymbol{w}_h) - s_h^{\mathrm{n}}((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)), \tag{63.28}$$

for all $\boldsymbol{v}_h \in \boldsymbol{V}_h$, we have

$$\begin{aligned}
\mu^{-1}\|q_h\|_{L^2(D)}^2 = t_h((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)) &- a_h(\boldsymbol{v}_h,\boldsymbol{w}_h) - s_h^{\mathrm{u}}(\boldsymbol{v}_h,\boldsymbol{w}_h) \\
&+ \sum_{F\in\mathcal{F}_h^{\circ}\cup\mathcal{F}_h^{\mathrm{n}}} (\llbracket q_h\rrbracket \boldsymbol{n}_F, \{\boldsymbol{w}_h - \boldsymbol{w}_{q_h}\})_{\boldsymbol{L}^2(F)} - s_h^{\mathrm{n}}((\boldsymbol{v}_h,q_h),(\boldsymbol{w}_h,0)).
\end{aligned}$$

Let $\mathfrak{T}_1, \mathfrak{T}_2, \mathfrak{T}_3, \mathfrak{T}_4, \mathfrak{T}_5$ denote the five terms on the right-hand side. We have

$$\begin{aligned}
&|\mathfrak{T}_1| \leq \mathbb{S}\|(\boldsymbol{w}_h,0)\|_{Y_h} \leq c\,\mathbb{S}\mu^{-\frac12}\|q_h\|_{L^2(D)}, \\
&|\mathfrak{T}_2 + \mathfrak{T}_3| \leq c\,\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}\|\boldsymbol{w}_h\|_{\boldsymbol{V}_h} \leq c'\,\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}\mu^{-\frac12}\|q_h\|_{L^2(D)},
\end{aligned}$$

where we used a discrete trace inequality to bound the contributions of n_h to a_h and the bound (63.25) on $\boldsymbol{w}_h$. Moreover, distinguishing the contribution of the interfaces and of the Neumann boundary faces, we write $\mathfrak{T}_4 := \mathfrak{T}_4^\circ + \mathfrak{T}_4^{\rm n}$ with obvious notation. The Cauchy–Schwarz inequality along with the estimate $\|\boldsymbol{w}_h - \boldsymbol{w}_{q_h}\|_{\boldsymbol{L}^2(F)} \le c h_F^{\frac12} |\boldsymbol{w}_{q_h}|_{\boldsymbol{H}^1(\mathcal{T}_F)}$ implies that

$$|\mathfrak{T}_4^\circ| \le c\,|q_h|_{S^{\rm p}} \|\boldsymbol{w}_h\|_{\boldsymbol{V}_h} \le c'\,|q_h|_{S^{\rm p}} \mu^{-\frac12} \|q_h\|_{L^2(D)},$$

whereas for the Neumann faces, we obtain

$$\mathfrak{T}_4^{\rm n} = \sum_{F\in\mathcal{F}_h^{\rm n}} (\mathbb{r}(\boldsymbol{v}_h,q_h)\boldsymbol{n}, \boldsymbol{w}_h - \boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)} + (\mathbb{s}(\boldsymbol{v}_h)\boldsymbol{n}, \boldsymbol{w}_h - \boldsymbol{w}_{q_h})_{\boldsymbol{L}^2(F)},$$

so that $|\mathfrak{T}_4^{\rm n}| \le c(|(\boldsymbol{v}_h,q_h|_{S^{\rm n}} + \|\boldsymbol{v}_h\|_{\boldsymbol{V}_h})\mu^{-\frac12}\|q_h\|_{L^2(D)}$. Since $s_h^{\rm n}$ is symmetric positive definite, we have $|\mathfrak{T}_5| \le |(\boldsymbol{v}_h,q_h)|_{S^{\rm n}} |(\boldsymbol{w}_h,0)|_{S^{\rm n}}$, and using (63.25) we infer that

$$|\mathfrak{T}_5| \le c\mu^{-1} |(\boldsymbol{v}_h,q_h)|_{S^{\rm n}} \|q\|_{L^2(D)}.$$

Putting everything together, we infer that $\|(\boldsymbol{v}_h,q_h)\|_{Y_h}^2 \le c(\mathbb{S}^2 + \mathbb{S}\|(\boldsymbol{v}_h,q_h)\|_{Y_h})$, and we conclude by invoking Young's inequality. □

Remark 63.11 (Inf-sup condition on b_h). Assume that $\partial D = \partial D_{\rm d}$. One can show that there is $\beta_0 > 0$ s.t. for all $q_h \in Q_h$, all $h \in \mathcal{H}$, and all $\mu > 0$,

$$\beta_0\, \mu^{-\frac12} \|q_h\|_{L^2(D)} \le \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h} \frac{|b_h(\boldsymbol{v}_h,q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}_h}} + |q_h|_{S^{\rm p}}, \tag{63.29}$$

see Exercise 63.6. An alternative proof of (63.24) using the inf-sup condition (63.29) is also given in Exercise 63.6. □

63.2.3 Error analysis

The error analysis proceeds as for the CIP method. We assume for that the solution to (62.5) is in the space $Y_{\rm s}$ defined in (63.14), and we set $Y_\sharp := Y_{\rm s} + Y_h$. We denote by $\|\cdot\|_{Y_\sharp}$ the natural extension of the norm $\|\cdot\|_{Y_h}$ defined in (63.23c) to $Y_\sharp$ (one readily verifies that this extension indeed defines a norm). We denote by $t_\sharp$ the natural extension of the bilinear form t_h defined in (63.21) to $Y_\sharp \times Y_h$. These extensions are meaningful since $r > \frac12$.

Theorem 63.12 (Error estimate). *Let $(\boldsymbol{u},p)$ solve (62.5) and assume that $(\boldsymbol{u},p) \in Y_{\rm s}$ with $Y_{\rm s}$ defined in (63.14). Let $(\boldsymbol{u}_h,p_h) \in Y_h$ solve (62.11) with the stabilizing bilinear forms $s_h^{\rm p}, s_h^{\rm n}$ defined in (62.10) and $s_h^{\rm u}$ defined in (63.20). Assume that $\varpi^{\rm u} > n_\partial c_{\rm dt}^2$.* (i) *There is c such that for all $h \in \mathcal{H}$ and all $\mu > 0$,*

$$\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)\|_{Y_\sharp} \le c \inf_{(\boldsymbol{v}_h,q_h)\in Y_h} \|(u-\boldsymbol{v}_h, p-q_h)\|_{Y_\sharp}. \tag{63.30}$$

(ii) *The following holds true for all $\tau \in (\frac12, \min(k_u, k_p+1)]$:*

$$\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)\|_{Y_\sharp} \le c \left(\sum_{K\in\mathcal{T}_h} \mu h_K^{2\tau} |\boldsymbol{u}|_{\boldsymbol{H}^{1+\tau}(K)}^2 + \frac{1}{\mu} h_K^{2\tau} |p|_{H^\tau(K)}^2 \right)^{\frac12}. \tag{63.31}$$

Proof. (i) The error estimate (63.30) follows from Lemma 62.4. Note in particular that to establish the Galerkin orthogonality property, we reason as in the proof of Lemma 38.2 to show that $(\boldsymbol{f},\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\nabla\cdot(\mathbb{r}(\boldsymbol{u},p)),\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = a_h(\boldsymbol{u},\boldsymbol{w}_h) - n_h(\boldsymbol{u},\boldsymbol{w}_h) + b_h(\boldsymbol{w}_h,p)$ (see Exercise 63.3).
(ii) The estimate (63.31) follows from (63.30) by using $\boldsymbol{v}_h := \boldsymbol{\mathcal{I}}_h^{\rm b}(\boldsymbol{u})$, $q_h := \mathcal{I}_h^{\rm b}(p)$ in the infima, and by invoking the approximation properties of the L^2-orthogonal projections. □

Exercises

Exercise 63.1 (Coercivity, CIP). Prove Lemma 63.2. (*Hint*: see the proofs of Lemma 37.2 and Lemma 37.3.)

Exercise 63.2 (Inf-sup condition on b, CIP). Prove the inf-sup condition (63.13) on b. Here, we do not assume that Q_h is H^1-conforming, that is, the pressure space is either $P^{\mathrm{g}}_{k_p}(\mathcal{T}_h)$ or $P^{\mathrm{b}}_{k_p}(\mathcal{T}_h)$. (*Hint*: use the identities for $\mu^{-1}h^2\|\nabla_h q_h\|^2_{\boldsymbol{L}^2(D)}$ and $\mu^{-1}\|q_h\|^2_{L^2(D)}$ from the proof of Lemma 63.3.)

Exercise 63.3 (Galerkin orthogonality, dG). Prove the Galerkin orthogonality for the stabilized dG formulation from §63.2, i.e., $t_h((\boldsymbol{u},p),(\boldsymbol{w}_h,r_h)) = \ell_h(\boldsymbol{w}_h,r_h)$ for all $(\boldsymbol{w}_h,r_h)\in Y_h$.

Exercise 63.4 (Integration by parts for b_h, dG). Let b_h be defined in (63.19). Prove the identity (63.27). (*Hint*: $[\![ab]\!] = \{a\}[\![b]\!] + [\![a]\!]\{b\}$ at all the interfaces.)

Exercise 63.5 (dG fluxes). Derive local formulations of the discrete problem using the fluxes from Remark 63.7. (*Hint*: proceed as in §38.4.)

Exercise 63.6 (Inf-sup conditions, dG). Assume that $\partial D = \partial D_{\mathrm{d}}$. (i) Prove the inf-sup condition (63.29) on b_h. (*Hint*: use (63.26).) (ii) Using the inf-sup condition on b_h, prove again the inf-sup condition on t_h. (*Hint*: use the identity (63.28).)

Chapter 64

Bochner integration

In Part XIII, composed of Chapters 64 to 71, we start the study of time-dependent PDEs. We focus on parabolic equations where the differential operator in space enjoys a coercivity property. We introduce suitable functional spaces for the weak formulation, and we establish its well-posedness by invoking either coercivity arguments or inf-sup conditions. Then we address the discretization in space and in time. We investigate the method of lines where the space discretization is done first. This leads to a finite system of ordinary differential equations which is then discretized by using some time-stepping technique. Prototypical examples include the Euler schemes (implicit or explicit), second-order schemes such as BDF2 and Crank–Nicolson, and higher-order schemes based on a space-time weak formulation leading to discontinuous Galerkin and continuous Petrov–Galerkin approximations, which are also called implicit Runge–Kutta (IRK) in the literature.

The goal of this chapter is to introduce a mathematical setting to formulate parabolic problems in some weak form. The viewpoint we are going to develop is to consider functions defined on a bounded time interval, say J, with values in some Banach (or Hilbert) space composed of functions defined on the space domain, say D. The key notions we develop in this chapter are the Bochner integral and the weak time derivative of functions that are Bochner integrable.

64.1 Bochner integral

We give in this section a brief overview of the Bochner integral theory. This theory is useful to deal with time-dependent functions with values in Banach spaces. Let J be a nonempty, bounded, open set in $\mathbb{R}$. Let V be a Banach space (real or complex). The main example we have in mind is $J := (0,T)$, $T > 0$, and V is composed of functions defined on some Lipschitz domain D in $\mathbb{R}^d$. The material is adapted from Kufner et al. [207, §2.19].

64.1.1 Strong measurability and Bochner integrability

Definition 64.1 (Simple functions). *We say that $f : J \to V$ is a* simple function *if there exist $m \in \mathbb{N}$, a finite collection of vectors $\{v_k\}_{k\in\{1:m\}}$ in V, and disjoint (Lebesgue) measurable subsets $\{A_k\}_{k\in\{1:m\}}$ in J, such that $f(t) = \sum_{k\in\{1:m\}} v_k \mathbb{1}_{A_k}(t)$ for all $t \in J$. The* Bochner integral *of a simple function is defined by*

$$\int_J f(t)\mathrm{d}t := \sum_{k\in\{1:m\}} v_k|A_k| \in V. \tag{64.1}$$

Lemma 64.2 (Bound on Bochner integral). *We have* $\| \int_J f(t)\mathrm{d}t\|_V \le \int_J \|f(t)\|_V \mathrm{d}t$ *for every simple function* f*.*

Proof. (64.1) and the triangle inequality in V imply that $\| \int_J f(t)\mathrm{d}t\|_V \le \sum_{k\in\{1:m\}} \|v_k\|_V |A_k| = \sum_{k\in\{1:m\}} \int_J \|v_k\|_V \mathbb{1}_{A_k}(t)\mathrm{d}t = \int_J \|f(t)\|_V \mathrm{d}t$, where the last equality results from the identity $\|f(t)\|_V = \sum_{k\in\{1:m\}} \|v_k\|_V \mathbb{1}_{A_k}(t)$ which is a consequence of the sets $\{A_k\}_{k\in\{1:m\}}$ being disjoint. □

Definition 64.3 (Strong measurability). *We say that* $f : J \to V$ *is* strongly measurable *if there is a countable sequence of simple functions* $(f_n)_{n\in\mathbb{N}}$ *such that* $\lim_{n\to\infty} \|f(t) - f_n(t)\|_V = 0$ *for a.e.* t *in* J*. Thus, a strongly measurable function is the limit (in the norm of* V*) of simple functions for a.e.* $t \in J$*.*

Let V' be the dual space of V. Recall that in the complex case, V' is composed of antilinear forms (see Definition A.11).

Theorem 64.4 (Pettis). *A function* $f : J \to V$ *is strongly measurable if and only if it satisfies the following two properties:*

(i) f *is weakly measurable: the function* $\langle v', f\rangle_{V',V} : J \ni t \mapsto \langle v', f(t)\rangle_{V',V} \in \mathbb{R}$ *(or* $\mathbb{C}$*) is Lebesgue measurable for all* $v' \in V'$*.*

(ii) f *is almost separably valued: There exists* $E \subset J$ *of zero measure s.t.* $f(J\backslash E)$ *is separable (i.e.,* $f(J\backslash E)$ *contains a countable dense subset).*

Proof. See Pettis measurability theorem in Diestel [107, Chap. IV, p. 25]. See also Showalter [257, Thm. 1.1, Chap. III]. □

Example 64.5 ($V = L^\infty(0,1)$). Let $J := (0,1)$ and $f : J \to V := L^\infty(D)$ with $D := (0,1)$ be defined by $f(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$, i.e., $f(t)(x) := 1$ if $x \in (0,t)$ and $f(t)(x) := 0$ otherwise. Then f is not almost separably valued. Let indeed $E \subset J$ be a subset of zero measure and F be a countable subset of $J\backslash E$. Notice that $|E\cup F| = 0$ so that $|J\backslash(E\cup F)| = 1$. Hence, $J\backslash(E\cup F)$ is not empty. Let $t \in J\backslash(E\cup F) \subset J\backslash E$. For all $s \in F$, we have $|f(t) - f(s)| = \mathbb{1}_{(\min(s,t),\max(s,t))}$. Since s cannot be equal to t, we have $|(\min(s,t),\max(s,t))| > 0$. This implies that $\|f(t)-f(s)\|_{L^\infty(D)} = 1$, so that there cannot exist any sequence $(s_n)_{n\in\mathbb{N}}$ in F so that $\|f(t) - f(s_n)\|_{L^\infty(D)}$ converges to 0. Hence, $f(F)$ cannot be dense in $f(J\backslash E)$. In conclusion, f is not strongly measurable. □

Example 64.6 ($V = L^2(0,1)$). Let $J := (0,1)$ and $g : J \to V := L^2(D)$ with $D := (0,1)$ be defined by $g(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$. Observe first that g is almost separably valued since $L^2(D)$ is separable. Identifying $(L^2(D))'$ with $L^2(D)$, we also have $\langle w, g(t)\rangle_{(L^2)',L^2} = (w, g(t))_{L^2(D)} = \int_D w(x)g(t)(x)\,\mathrm{d}x = \int_0^t w(x)\,\mathrm{d}x$ for every $w \in (L^2(D))' = L^2(D)$. The function $J \ni t \mapsto \int_0^t w(x)\,\mathrm{d}x$ is measurable since it is continuous. Hence, g is weakly measurable. In conclusion, g is strongly measurable. □

Lemma 64.7 (Measurability of norm). *Let* $f : J \to V$ *be strongly measurable. Then the map* $J \ni t \mapsto \|f(t)\|_V \in \mathbb{R}$ *is Lebesgue measurable.*

Proof. See Kufner et al. [207, Lem. 2.19.2]. □

Example 64.8 (Semi-discrete function). Let $V_h \subset V$ be a finite-dimensional subspace. Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h with $I := \dim(V_h)$, and let $\{\psi_i\}_{i\in\{1:I\}}$ be functions in $L^1(J;\mathbb{R})$. The function $f : J \to V$ such that $f(t) := \sum_{i\in\{1:I\}} \psi_i(t)\varphi_i$ is strongly measurable (see Exercise 64.1). Functions of this form play a central role in the semi-discretization in space of the model problem (65.1); see Chapter 66. □

Definition 64.9 (Bochner integrability). *We say that $f : J \to V$ is* Bochner integrable *if there exists a countable sequence of simple functions $(f_n)_{n\in\mathbb{N}}$ s.t. $\lim_{n\to\infty} \|f(t) - f_n(t)\|_V = 0$ for a.e. t in J (i.e., f is strongly measurable), and $\lim_{n\to\infty} \int_J \|f(t) - f_n(t)\|_V \mathrm{d}t = 0$.*

Lemma 64.10 (Limit of integrals). *Let $f : J \to V$ be a Bochner integrable function and $(f_n)_{n\in\mathbb{N}}$ be a sequence of simple functions as in Definition 64.9. Then $(\int_J f_n(t)\mathrm{d}t)_{n\in\mathbb{N}}$ converges in V. Moreover, if $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are two sequences with the above properties, then their integrals have the same limit.*

Proof. See Exercise 64.2. □

Definition 64.11 (Bochner integral). *Let $f : J \to V$ be a Bochner integrable function. The* Bochner integral *of f is defined as $\lim_{n\to\infty} \int_J f_n(t)\mathrm{d}t$ (the convergence occurs in the norm of V), where $(f_n)_{n\in\mathbb{N}}$ is any countable sequence of simple functions as in Definition 64.9.*

Theorem 64.12 (Bochner). *A strongly measurable function $f : J \to V$ is Bochner integrable if and only if $\int_J \|f(t)\|_V \mathrm{d}t < \infty$.*

Proof. See Kufner et al. [207, Thm. 2.19.8], Diestel [107, p. 26]. □

In this book, we are only going to manipulate strongly measurable functions. Theorem 64.12 then says that to verify that f is Bochner integrable, it suffices to verify that $J \ni t \mapsto \|f(t)\|_V \in \mathbb{R}$ is in $L^1(J;\mathbb{R})$.

Example 64.13 ($V = L^2(0,1)$). Let g be the function from Example 64.6, i.e., $J := (0,1)$, $V := L^2(D)$ with $D := (0,1)$, and $g(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$. We have $\|g(t)\|_V^2 = \int_0^1 \mathbb{1}_{(0,t)}^2 \,\mathrm{d}x = t$. Hence, $\int_J \|g(t)\|_V \mathrm{d}t = \int_0^1 t^{\frac{1}{2}} \mathrm{d}t = \frac{2}{3}$. This shows that g is Bochner integrable. □

64.1.2 Main properties

In this section, we state some useful properties of the Bochner integral.

Corollary 64.14 (Linear map). *Let V, W be Banach spaces and consider a bounded linear operator $K \in \mathcal{L}(V;W)$. Let $f : J \to V$ be a Bochner integrable function, and define the function $K(f) : J \to W$ such that $(K(f))(t) := K(f(t))$ for a.e. $t \in J$. Then $K(f)$ is a Bochner integrable function, and $\int_J (K(f))(t)\mathrm{d}t = K\left(\int_J f(t)\mathrm{d}t\right)$.*

Proof. See Kufner et al. [207, Cor. 2.19.11]. □

Example 64.15 (Linear forms). Let $f : J \to V$ be a Bochner integrable function. Let $\psi \in V'$. Then $\int_J \langle \psi, f(t)\rangle_{V',V} \mathrm{d}t = \langle \psi, \int_J f(t)\mathrm{d}t\rangle_{V',V}$. □

Example 64.16 (Embedding). Let $f : J \to V$ be a Bochner integrable function. Let L be another Banach space such that $V \hookrightarrow L$. Denote by $K_{V\to L} : V \to L$ the canonical embedding. Then $K_{V\to L}(\int_J f(t)\mathrm{d}t) = \int_J K_{V\to L}(f(t))\mathrm{d}t$. This means that the L-valued and the V-valued integrals of f can be identified, which we are going to do systematically. □

Definition 64.17 ($L^p(J;V)$). *Let $p \in [1,\infty]$. We call* Bochner space *$L^p(J;V)$ the space composed of the functions $v : J \to V$ that are strongly measurable and such that the following norm is finite:*

$$\|v\|_{L^p(J;V)} := \begin{cases} \left(\int_J \|v(t)\|_V^p \mathrm{d}t\right)^{\frac{1}{p}} & \text{if } p \in [1,\infty), \\ \operatorname{ess\,sup}_{t\in J} \|v(t)\|_V & \text{if } p = \infty. \end{cases} \tag{64.2}$$

One can verify that the following properties hold true: (i) $\| \int_J v(t) \mathrm{d}t \|_V \leq \int_J \|v(t)\|_V \mathrm{d}t = \|v\|_{L^1(J;V)}$ for every Bochner integrable function $v : J \to V$; (ii) $L^p(J;V) \hookrightarrow L^1(J;V)$. See Exercise 64.3.

Lemma 64.18 (Lebesgue's dominated convergence). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $L^1(J;V)$. Assume that $(f_n(t))_{n\in\mathbb{N}}$ converges to $f(t)$ in V for a.e. $t \in J$ and there is $g \in L^1(J;\mathbb{R})$ such that $\|f_n(t)\|_V \leq g(t)$ for a.e. $t \in J$. Then $f \in L^1(J;V)$ and $(f_n)_{n\in\mathbb{N}}$ converges to f in $L^1(J;V)$.*

Proof. See Exercise 64.3. □

Theorem 64.19 (Banach space). *Let V be a Banach space. Then $L^p(J;V)$ is a Banach space for all $p \in [1,\infty]$.*

Proof. See Kufner et al. [207, Thm. 2.20.4]. □

Theorem 64.20 (Dual space). *Let V be a reflexive Banach space and $p \in [1,\infty)$.* (i) *The dual space of $L^p(J;V)$ is isometrically isomorphic to $L^{p'}(J;V')$, $\frac{1}{p} + \frac{1}{p'} = 1$.* (ii) *$L^p(J;V)$ is reflexive for all $p \in (1,\infty)$.*

Proof. This is Theorem 3.2 in Bochner and Taylor [36]. See also Theorem 2.22.3 and the remarks on page 125 in Kufner et al. [207]. □

Example 64.21 ($L^p(J;L^p(D)) = L^p(J\times D)$). Let D be a nonempty open subset of $\mathbb{R}^d$. Recall that $L^p(D)$ is a Banach space for all $p \in [1,\infty]$. Using Fubini's theorem (Theorem 1.47), one can identify the Bochner space $L^p(J;L^p(D))$ with the Lebesgue space $L^p(J\times D)$ for all $p \in [1,\infty)$. □

Example 64.22 ($L^q((0,1);L^2(0,1))$). Let g be the function from Example 64.6, i.e., $J := (0,1)$, $V := L^2(D)$ with $D := (0,1)$, and $g(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$. We know that g is strongly measurable and we have $\|g(t)\|_V^2 = \int_0^1 \mathbb{1}_{(0,t)}^2 \,\mathrm{d}x = t$. Let $q \in [1,\infty)$. Then $\|g\|_{L^q(J;V)} = (\int_0^1 t^{\frac{q}{2}} \mathrm{d}t)^{\frac{1}{q}} = (\frac{2}{q+2})^{\frac{1}{q}}$. Hence, $g \in L^q((0,1);L^2(0,1))$. □

Theorem 64.23 (Density). *Let V be a Banach space and $p \in [1,\infty)$. Then simple functions are dense in $L^p(J;V)$.*

Proof. Let $f \in L^p(J;V)$. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of simple functions converging to f. For all $n \in \mathbb{N}$, we set $g_n := f_n \mathbb{1}_{B_n}$, where $B_n := \{t \in J \mid \|f_n(t)\|_V \leq 2\|f(t)\|_V\}$. Note that $(g_n)_{n\in\mathbb{N}}$ is a sequence of simple functions. This definition implies that $\|g_n(t) - f(t)\|_V \to 0$ for a.e. t in J as $n \to \infty$. Moreover, $\sup_{n\in\mathbb{N}} \|g_n(t) - f(t)\|_V \leq \sup_{n\in\mathbb{N}} \|g_n(t)\|_V + \|f(t)\|_V \leq 3\|f(t)\|_V$. Hence (recalling that $p < \infty$), Lebesgue's dominated convergence theorem applied to $\|g_n - f\|_V^p$ in $L^1(J;\mathbb{R})$ implies that $\int_J \|g_n(t) - f(t)\|_V^p \mathrm{d}t \to 0$ as $n \to \infty$, which proves the assertion. □

Remark 64.24 (Tensor products). Let $L^p(J) := L^p(J;\mathbb{R})$ if V is a real vector space and $L^p(J) := L^p(J;\mathbb{C})$ if V is a complex vector space. The vector space $L^p(J) \otimes V$ is by definition composed of all the functions f in $L^p(J;V)$ such that there exists a finite collection of vectors $\{v_k\}_{k\in\{1:m\}}$ in V and functions $\{\phi_k\}_{k\in\{1:m\}}$ in $L^p(J)$, for some $m \in \mathbb{N}$, such that $f(t) := \sum_{k\in\{1:m\}} v_k \phi_k(t)$. Simple functions are members of $L^p(J) \otimes V$. Hence, Theorem 64.23 implies that $L^p(J) \otimes V$ is dense in $L^p(J;V)$. □

64.2 Weak time derivative

In this section, we study the important notion of weak time derivative in Bochner spaces. We also show that pointwise values in time are meaningful for functions having an integrable weak time derivative. These two notions are fundamental to the weak formulation of parabolic problems, as we shall see in the next chapter. In the entire section, we take $J := (0,T)$ with $T > 0$.

64.2.1 Strong and weak time derivatives

Definition 64.25 (Continuity). *A function $f : J \to V$ is said to be continuous at $t \in J$ (in the norm topology, or strong topology) if for every sequence $(t_n)_{n\in\mathbb{N}}$ that converges to t in J, the sequence $(f(t_n))_{n_\mathbb{N}}$ converges to $f(t)$ in V, and f is said to be continuous if it is continuous at every point in J.*

We denote by $C^0(J;V)$ the space composed of the functions $v : J \to V$ that are continuous, and we set $C^0(\overline{J};V) := C^0(J;V) \cap L^\infty(J;V)$. This space, equipped with the norm $\|v\|_{C^0(\overline{J};V)} := \sup_{t\in\overline{J}} \|v(t)\|_V$, is a Banach space.

Definition 64.26 (Strong time derivative). *Let V be a Banach space. Let $f : J \to V$. Assume that f is continuous in a neighborhood of $t \in J$. We say that f is (strongly) differentiable at t if the ratio $\frac{f(t+\tau)-f(t)}{\tau}$ converges strongly in V as $\tau \to 0$. The limit is then denoted by $\partial_t f(t) \in V$.*

Theorem 64.27 (Lebesgue's differentiation). *Let V be a Banach space and $f \in L^1(J;V)$. Let $F(t) := \int_0^t f(\xi)\,\mathrm{d}\xi$ for all $t \in J$. Then F is strongly differentiable for a.e. $t \in J$ and $\partial_t F(t) = f(t)$ for a.e. $t \in J$.*

Proof. Let $t \in J$ and τ be small enough so that $t + \tau \in J$. Then $\frac{F(t+\tau)-F(t)}{\tau} = \frac{1}{\tau}\int_t^{t+\tau} f(\xi)\,\mathrm{d}\xi$. Denoting $R_\tau(t) := f(t) - \frac{1}{\tau}\int_t^{t+\tau} f(\xi)\,\mathrm{d}\xi$, we need to establish that $\lim_{\tau\to 0}\|R_\tau(t)\|_V = 0$ for a.e. $t \in J$. Since f is strongly measurable, we infer from Pettis theorem (Theorem 64.4) that f is almost separably valued, i.e., there is a subset $E \subset J$ of zero measure such that $f(J\backslash E)$ is separable. Let $\{a_n \in V \mid n \in \mathbb{N}\}$ be a countable dense subset of $f(J\backslash E)$. By applying Lebesgue's differentiation theorem to the real-valued function $\|f - a_n\|_V$ (see Theorem 2.1), we infer that for all $n \in \mathbb{N}$, there is a subset S_n of J of zero measure s.t. $\|f(t) - a_n\|_V = \lim_{\tau\to 0}\frac{1}{\tau}\int_t^{t+\tau}\|f(\xi) - a_n\|_V\,\mathrm{d}\xi$ for all $t \in J\backslash S_n$. Since $\|R_\tau(t)\|_V \le \frac{1}{\tau}\int_t^{t+\tau}\|f(t) - a_n + a_n - f(\xi)\|_V\,\mathrm{d}\xi$, we infer that

$$\limsup_{\tau\to 0}\|R_\tau(t)\|_V \le 2\|f(t) - a_n\|_V, \quad \text{for a.e. } t \in J\backslash S_n.$$

Since the above inequality holds true for all $n \in \mathbb{N}$ and the set $\{a_n \in V \mid n \in \mathbb{N}\}$ is dense in $f(J\backslash E)$, we conclude that $\limsup_{\tau\to 0}\|R_\tau(t)\|_V = 0$ for all $t \in J\backslash(E\bigcup_{n\in\mathbb{N}} S_n)$. By the subadditivity property of the Lebesgue measure (see page 2, Chapter 1), we have $|\bigcup_{n\in\mathbb{N}} S_n| = 0$. Hence, $\limsup_{\tau\to 0}\|R_\tau(t)\|_V = 0$ for a.e. $t \in J$. This proves the assertion. □

Corollary 64.28 (Vanishing integral in $L^1_{\mathrm{loc}}(J;V)$). *Let $L^1_{\mathrm{loc}}(J;V)$ be the space composed of the functions from J to V that are strongly measurable and are integrable over every subset that is compact in J. Let $f \in L^1_{\mathrm{loc}}(J;V)$ be s.t. $\int_J f(t)\phi(t)\mathrm{d}t = 0$ for all $\phi \in C_0^\infty(J;\mathbb{R})$. Then $f = 0$ for a.e. $t \in J$.*

Proof. Let $\tau \in J$ and let $t \in (\tau,T)$. Consider a sequence of functions $(\phi_n)_{n\in\mathbb{N}}$ in $C_0^\infty(J;\mathbb{R})$ such that $\phi_n(\xi) \in [0,1]$ for all $\xi \in J$, and $\phi_n \to \mathbb{1}_{(\tau,t)}$ a.e. in J. Then Lebesgue's dominated convergence theorem in $L^1(J;V)$ implies that $0 = \lim_{n\to\infty}\int_J \phi_n(\xi)f(\xi)\,\mathrm{d}\xi = \int_\tau^t f(\xi)\,\mathrm{d}\xi$ for a.e. $t \in (\tau,T)$. We conclude that $f(t) = 0$ for a.e. $t \in (\tau,T)$ by invoking Theorem 64.27 with $J := (\tau,T)$. This proves the assertion since τ can be arbitrarily close to 0. □

Definition 64.29 (Weak time derivative). *Let V be a Banach space. We say that the function $v \in L^1_{\mathrm{loc}}(J;V)$ has a* weak time derivative *if there is $w \in L^1_{\mathrm{loc}}(J;V)$ such that*

$$-\int_J \phi'(t)v(t)\mathrm{d}t = \int_J \phi(t)w(t)\mathrm{d}t, \quad \forall \phi \in C_0^\infty(J;\mathbb{R}). \tag{64.3}$$

We denote by $\partial_t v := w$ the weak time derivative of v. When the context is unambiguous, we simply say that the function v is weakly differentiable and call $\partial_t v$ its weak derivative.

Lemma 64.30 (Constants). *Let V be a Banach space and $f \in L^1_{\mathrm{loc}}(J;V)$. Assume that f is weakly differentiable and $\partial_t f = 0$. Then there is $a \in V$ such that $f(t) = a$ for a.e. $t \in J$.*

Proof. See Exercise 64.5. □

Theorem 64.31 (Fundamental theorem of calculus). *Let V be a Banach space, $f \in L^1(J;V)$, and $g \in L^1(J;V)$. Then f is weakly differentiable with $\partial_t f = g$ iff there is $a \in V$ s.t. $f(t) = a + \int_0^t g(\xi)\,\mathrm{d}\xi$ for a.e. $t \in J$.*

Proof. (1) Assume that $f(t) = a + \int_0^t g(\xi)\,\mathrm{d}\xi$ for a.e. $t \in J$ and let us show that g is the weak derivative of f. Let $\phi \in C_0^\infty(J;\mathbb{R})$. Fubini's theorem gives

$$\begin{aligned}-\int_J \phi'(t)f(t)\mathrm{d}t &= \lim_{\tau\to 0} -\int_J \frac{\phi(t+\tau)-\phi(t)}{\tau} f(t)\mathrm{d}t = \lim_{\tau\to 0}\int_J \phi(t)\frac{f(t)-f(t-\tau)}{\tau}\mathrm{d}t \\ &= \lim_{\tau\to 0}\int_J \frac{\phi(t)}{\tau}\int_{t-\tau}^t g(\xi)\,\mathrm{d}\xi\mathrm{d}t = \lim_{\tau\to 0}\int_J g(\xi)\int_J \phi(t)\frac{\mathbb{1}_{(t-\tau,t)}(\xi)}{\tau}\mathrm{d}t\,\mathrm{d}\xi.\end{aligned}$$

Since $\int_J \phi(t)\frac{\mathbb{1}_{(t-\tau,t)}(\xi)}{\tau}\mathrm{d}t = \frac{1}{\tau}\int_\xi^{\xi+\tau}\phi(t)\mathrm{d}t \to \phi(\xi)$ uniformly with respect to $\xi \in J$ as $\tau \to 0$, and $|\int_J \phi(t)\frac{\mathbb{1}_{(t-\tau,t)}(\xi)}{\tau}\mathrm{d}t| \le \|\phi\|_{C^0(\overline{J})}$ for every $\xi \in J$, Lebesgue's dominated convergence theorem in $L^1(J;V)$ implies that $g(\xi)\int_J \phi(t)\frac{\mathbb{1}_{(t-\tau,t)}(\xi)}{\tau}\mathrm{d}t \to \phi(\xi)g(\xi)$ in $L^1(J;V)$ as $\tau \to 0$. The above computation shows that

$$\int_J \phi'(t)f(t)\mathrm{d}t = -\int_J \phi(t)g(t)\mathrm{d}t, \qquad \forall \phi \in C_0^\infty(J;\mathbb{R}).$$

This proves that g is the weak derivative of f.
(2) Conversely, let us assume that f is weakly differentiable with $\partial_t f = g \in L^1(J;V)$. Let us set $\widetilde{f}(t) := \int_0^t g(\xi)\,\mathrm{d}\xi$. The above argument shows that $\widetilde{f}$ is weakly differentiable and $\partial_t \widetilde{f} = g$. We invoke Lemma 64.30 to conclude that there is $a \in V$ s.t. $f(t) - \widetilde{f}(t) = a$. □

Corollary 64.32 (Strong vs. weak). *Let V be a Banach space and $f \in L^1(J;V)$ be weakly differentiable, i.e., $\partial_t f \in L^1_{\mathrm{loc}}(J;V)$. Then f is strongly differentiable a.e. in J, and its strong and weak derivatives coincide.*

Proof. The assertion is a simple consequence of Theorem 64.31 and Lebesgue's differentiation theorem. □

Proposition 64.33 (Characterization). *Let V be a Banach space. Let $f, g \in L^1_{\mathrm{loc}}(J;V)$. Then f is weakly differentiable with $\partial_t f = g$ if and only if the map $J \ni t \mapsto \langle v', f(t)\rangle_{V',V} \in \mathbb{R}$ (or $\mathbb{C}$) is weakly differentiable for all $v' \in V'$, and $\partial_t\langle v', f\rangle_{V',V} = \langle v', g\rangle_{V',V}$ a.e. in J.*

Proof. Let f be weakly differentiable with $\partial_t f = g$. Then $-\int_J \phi'(t)f(t)\mathrm{d}t = \int_J \phi(t)g(t)\mathrm{d}t$ for all $\phi \in C_0^\infty(J;\mathbb{R})$. Using twice Corollary 64.14, we infer that

$$\int_J \phi'(t)\langle v', f(t)\rangle_{V',V}\mathrm{d}t = \langle v', \int_J \phi'(t)f(t)\mathrm{d}t\rangle_{V',V}$$
$$= -\langle v', \int_J \phi(t)g(t)\mathrm{d}t\rangle_{V',V} = -\int_J \phi(t)\langle v', g(t)\rangle_{V',V}\mathrm{d}t,$$

for all $v' \in V'$, which means that the map $J \ni t \mapsto \langle v', f(t)\rangle_{V',V}$ is weakly differentiable, and $\partial_t\langle v', f\rangle_{V',V} = \langle v', g\rangle_{V',V}$. Conversely, if $\partial_t\langle v', f\rangle_{V',V} = \langle v', g\rangle_{V',V}$ for all $v' \in V'$, then as above, we have $\langle v', \int_J \phi'(t)f(t)\mathrm{d}t\rangle_{V',V} = -\langle v', \int_J \phi(t)g(t)\mathrm{d}t\rangle_{V',V}$ for all $\phi \in C_0^\infty(J;\mathbb{R})$ and all $v' \in V'$. This proves that $\int_J \phi'(t)f(t)\mathrm{d}t = -\int_J \phi(t)g(t)\mathrm{d}t$ for all $\phi \in C_0^\infty(J;\mathbb{R})$, i.e., f is weakly differentiable with $\partial_t f = g$. □

Lemma 64.34 (Linear map). *Let V, W be two Banach spaces and let $K \in \mathcal{L}(V;W)$. Then for all weakly differentiable v in $L^1_{\mathrm{loc}}(J;V)$, $K(v)$ is in $L^1_{\mathrm{loc}}(J;W)$ and is weakly differentiable, and we have $K(\partial_t v) = \partial_t(K(v))$ in $L^1_{\mathrm{loc}}(J;W)$.*

Proof. See Exercise 64.6. □

64.2.2 Functional spaces with weak time derivative

Let V, W be two Banach spaces with continuous embedding $V \hookrightarrow W$ and canonical injection $K_{V\to W}$. We will often be in the situation where we have a function $v \in L^1_{\mathrm{loc}}(J;V)$ s.t. $\partial_t(K_{V\to W}(v)) \in L^1_{\mathrm{loc}}(J;W)$, recalling that $K_{V\to W}(v)(t) := K_{V\to W}(v(t))$. Then as in Example 64.16, we have $K_{V\to W}(\int_J \phi'(t)v(t)\mathrm{d}t) = \int_J \phi'(t)K_{V\to W}(v(t))\mathrm{d}t$, and we are going to abuse the notation by identifying $\partial_t(K_{V\to W}(v))$ and $\partial_t v$.

Definition 64.35 ($X^{p,q}(J;V,W)$). *Let $p, q \in [1,\infty]$. Let $V \hookrightarrow W$ be two Banach spaces with continuous embedding. We define*

$$X^{p,q}(J;V,W) := \{v \in L^p(J;V) \mid \partial_t v \in L^q(J;W)\}. \tag{64.4}$$

Consider the norm $\|\cdot\|_{X^{p,q}(J;V,W)} := \|\cdot\|_{L^p(J;V)} + \iota_{W,V}^{-1} T^{1+\frac{1}{p}-\frac{1}{q}}\|\partial_t\cdot\|_{L^q(J;W)}$ where $\iota_{W,V} := \sup_{v\in V}\frac{\|v\|_W}{\|v\|_V} = \|K_{V\to W}\|_{\mathcal{L}(V;W)}$. Notice that the two terms composing the norm $\|\cdot\|_{X^{p,q}(J;V,W)}$ are dimensionally consistent. One readily verifies that $X^{p,q}(J;V,W)$ is a Banach space when equipped with the above norm; see Exercise 64.7.

The following density result is of fundamental importance to study the properties of the space $X^{p,q}(J;V,W)$. Recall that $v \in C^\infty(\overline{J};V)$ if $v \in C^\infty(J;V)$ and v and all its derivatives have a continuous extension to $\overline{J}$.

Theorem 64.36 (Density). *$C^\infty(\overline{J};V)$ is dense in $X^{p,q}(J;V,W)$.*

Proof. This is Theorem 1 in Dautray and Lions [100, p. 473]. We propose here a slightly different proof, somewhat more direct, based on a shrinking mapping in the spirit of §23.1. Let us consider the kernel $\rho \in C^\infty(\mathbb{R};\mathbb{R})$ s.t. $\rho(s) := \eta e^{-\frac{1}{1-s^2}}$ if $s \in (-1,1)$, and $\rho := 0$ outside $(-1,1)$, where the real number η is chosen s.t. $\int_{\mathbb{R}} \rho(s)\,\mathrm{d}s = \int_{-1}^1 \rho(s)\,\mathrm{d}s = 1$. Let $\epsilon > 0$ and set $\varphi_\epsilon(t;s) = \frac{2t+(s+1)\epsilon T}{2(1+\epsilon)}$ for all $s \in (-1,1)$ and $t \in J$. Notice that $0 < \frac{t}{(1+\epsilon)} < \varphi_\epsilon(t;s) < \frac{t+\epsilon T}{(1+\epsilon)} < T$, so that $\varphi_\epsilon(t;s) \in J$ for all $s \in (-1,1)$ and $t \in J$. Let $u \in X^{p,q}(J;V,W)$ and consider the function $v_\epsilon : J \to V$ s.t. for all $t \in J$,

$$v_\epsilon(t) := \int_{-1}^1 \rho(s)u(\varphi_\epsilon(t;s))\,\mathrm{d}s.$$

Invoking standard arguments, we infer that $v_\epsilon \in C^\infty(\overline{J};V)$.

Let us prove that $v_\epsilon \to u$ in $L^p(J;V)$ as $\epsilon \to 0$. Since $u(t)-v_\epsilon(t) = \int_{-1}^1 \rho(s)(u(t)-u(\varphi_\epsilon(t;s)))\,\mathrm{d}s$, we have

$$\|u(t)-v_\epsilon(t)\|_V \le \|\rho\|_{L^\infty(\mathbb{R})} \int_{-1}^1 \|u(t)-u(\varphi_\epsilon(t;s))\|_V \,\mathrm{d}s.$$

(Note that $\|\rho\|_{L^\infty(\mathbb{R})} = \eta e^{-1}$.) Minkowsky's integral inequality implies that

$$\begin{aligned}\|u-v_\epsilon\|_{L^p(J;V)} &\le \|\rho\|_{L^\infty(\mathbb{R})} \left(\int_J \left(\int_{-1}^1 \|u(\varphi_\epsilon(t;s))-u(t)\|_V \,\mathrm{d}s \right)^p \mathrm{d}t \right)^{\frac{1}{p}} \\ &\le \|\rho\|_{L^\infty(\mathbb{R})} \int_{-1}^1 \left(\int_J \|u(\varphi_\epsilon(t;s))-u(t)\|_V^p \mathrm{d}t \right)^{\frac{1}{p}} \mathrm{d}s.\end{aligned}$$

Let $\mathcal{T}_{s,\epsilon} : L^p(J;V) \to L^p(J;V)$ be defined by $\mathcal{T}_{s,\epsilon}(u)(t) := u(\varphi_\epsilon(t;s))$ for all $s \in (-1,1)$ and $t \in J$. We can rewrite the above bound as $\|u-v_\epsilon\|_{L^p(J;V)} \le \|\rho\|_{L^\infty(\mathbb{R})} \int_{-1}^1 \|\mathcal{T}_{s,\epsilon}(u)-u\|_{L^p(J;V)} \,\mathrm{d}s$. We are going to show that $\mathcal{T}_{s,\epsilon}(u) \to u$ in $L^p(J;V)$ as $\epsilon \to 0$, uniformly w.r.t. $s \in (-1,1)$. Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of simple functions converging to u in $L^p(J;V)$. We have

$$\begin{aligned}\|\mathcal{T}_{s,\epsilon}(u)-u\|_{L^p(J;V)} \le{}& \|\mathcal{T}_{s,\epsilon}(u)-\mathcal{T}_{s,\epsilon}(u_n)\|_{L^p(J;V)} + \|\mathcal{T}_{s,\epsilon}(u_n)-u_n\|_{L^p(J;V)} \\ &+ \|u_n-u\|_{L^p(J;V)}.\end{aligned}$$

Making the change of variable $\varphi_\epsilon(t;s) \to z$ in the first term on the right-hand side, we obtain

$$\|\mathcal{T}_{s,\epsilon}(u)-u\|_{L^p(J;V)} \le \|\mathcal{T}_{s,\epsilon}(u_n)-u_n\|_{L^p(J;V)} + c_\epsilon \|u_n-u\|_{L^p(J;V)},$$

with $c_\epsilon := 1+(1+\epsilon)^{\frac{1}{p}}$. Since u_n is a simple function, we have $\lim_{\epsilon\to 0} \|\mathcal{T}_{s,\epsilon}(u_n)-u_n\|_{L^p(J;V)} = 0$ for all $s \in (-1,1)$. We infer that

$$\limsup_{\epsilon\to 0} \|\mathcal{T}_{s,\epsilon}(u)-u\|_{L^p(J;V)} \le 2\|u_n-u\|_{L^p(J;V)}.$$

The assertion follows readily since $\lim_{n\to\infty} \|u_n-u\|_{L^p(J;V)} = 0$. (Notice that it is essential to invoke the sequence of simple functions in the above argument.) Finally, using that $\partial_t v_\epsilon(t) := \frac{1}{1+\epsilon}\int_{-1}^1 \rho(s)\partial_t u(\varphi_\epsilon(t;s))\,\mathrm{d}s$, we can reason as above to prove that $\partial_t v_\epsilon \to \partial_t u$ in $L^q(J;V')$ as $\epsilon \to 0$. □

Lemma 64.37 (Embedding). (i) *$X^{p,q}(J;V,W)$ is continuously embedded in $C^{0,1-\frac{1}{q}}(\overline{J};W)$ if $q>1$ and in $C^0(\overline{J};W)$ if $q=1$.* (ii) *If V,W are Hilbert spaces, $X^{p,q}(J;V,W)$ is continuously embedded in $C^0(\overline{J},[V,W]_{\frac{1}{2},2})$ (see Definition* A.22 *for the interpolated space $[V,W]_{\frac{1}{2},2}$).*

Proof. See Exercise 64.8 for the proof of the first part and Lions and Magenes [220, Thm. 3.1] for the second part. □

Remark 64.38 (Continuous representative). Lemma 64.37(i) means that for every function u in $X^{p,q}(J;V,W)$, there exists a function $v \in C^{0,1-\frac{1}{q}}(\overline{J};W)$ s.t. $u(t) = v(t)$ for a.e. $t \in J$. It is then possible to replace u by its continuous representative v. We will systematically do this replacement in the rest of the book whenever a continuous embedding in a space of continuous functions is invoked; see also Remark 2.27. For instance, in Theorem 64.31 the function f is in $X^{1,1}(J;L^1(D),L^1(D))$, and denoting f^{c} the continuous representative of f, we have $f^{\mathrm{c}}(t) = f^{\mathrm{c}}(0) + \int_0^t \partial_t f(s)\,\mathrm{d}s$ for all $t \in J$. We will systematically abuse the notation and write f in lieu of f^{c}. □

Theorem 64.39 (Aubin–Lions–Simon). *Let $V \hookrightarrow M \hookrightarrow W$ be Banach spaces and assume that V is compactly embedded into M.* (i) *The embedding $X^{p,q}(J;V,W) \hookrightarrow L^p(J;M)$ is compact for all $p,q \in [1,\infty)$.* (ii) *The embedding $X^{\infty,q}(J;V,W) \hookrightarrow C^0(\overline{J};M)$ is compact for all $q > 1$.*

Proof. See Simon [261, Cor. 4, p. 85], Aubin [15, Thm. 2], Lions [219, Thm. I.5.1]; see also Amann [8, Thm. 5.1]. □

Let us now specialize the setting to separable Hilbert spaces (real or complex). Let $V \hookrightarrow L$ be two Hilbert spaces with continuous embedding and such that V is dense in L. We identify L with its dual L' so that

$$V \hookrightarrow L \equiv L' \hookrightarrow V', \tag{64.5}$$

where $L \equiv L'$ is dense in V'. The duality pairing $\langle \cdot, \cdot \rangle_{V',V}$ is viewed as an extension of the inner product in L, i.e., $\langle f, v \rangle_{V',V} = (f,v)_L$ for all $f \in L$ and all $v \in V$. Triples $(V, L \equiv L', V')$ with the above properties are often called *Gelfand triples* (see Brezis [52, Rmk. 3, p. 136]). Henceforth, we take $p = q := 2$ and $W := V'$ in Definition 64.35, and we omit the superscripts p and q, i.e., we write

$$X(J;V,V') := \{v \in L^2(J;V) \mid \partial_t v \in L^2(J;V')\}. \tag{64.6}$$

The following result justifies *integration by parts* with respect to time.

Lemma 64.40 (Time trace, integration by parts). *Let $(V, L \equiv L', V')$ be a Gelfand triple. Let $X(J;V,V')$ be defined in* (64.6)*. The following holds true:*

(i) $X(J;V,V') \hookrightarrow C^0(\overline{J};L)$.

(ii) *The map $X(J;V,V') \ni u \mapsto u(0) \in L$ is surjective.*

(iii) *The following integration by parts formula holds true: For all $v, w \in X(J;V,V')$,*

$$\int_J \langle \partial_t v(t), w(t) \rangle_{V',V} \mathrm{d}t = -\int_J \langle \partial_t w(t), v(t) \rangle_{V',V} \mathrm{d}t + (v(T), w(T))_L - (v(0), w(0))_L. \tag{64.7}$$

Proof. Item (i) is proved in Dautray and Lions [100, Thm. 2, p. 477], and Item (ii) is proved in Lions and Magenes [220, Thm. 3.2, p. 21] and [100, Rmk. 8, p. 523]. Let us prove Item (iii). Owing to Theorem 64.36, there are two sequences $(v_n)_{n\in\mathbb{N}}$, $(w_n)_{n\in\mathbb{N}}$ in $C^\infty(\overline{J};V)$ such that $v_n \to v$, $w_n \to w$ in $L^2(J;V)$, $v_n \to v$, $w_n \to w$ in $C^0(\overline{J};L)$, and $\partial_t v_n \to \partial_t v$, $\partial_t w_n \to \partial_t w$ in $L^2(J;V')$. Then we have

$$\int_J (\partial_t v_n, w_m)_L \mathrm{d}t = -\int_J (v_n, \partial_t w_m)_L \mathrm{d}t + (v_n(T), w_m(T))_L - (v_n(0), w_m(0))_L.$$

We conclude by passing to the limit in this identity. Indeed, we have

$$\lim_{n\to\infty} \int_J (\partial_t v_n, w_m)_L \mathrm{d}t = \lim_{n\to\infty} \int_J \langle \partial_t v_n, w_m \rangle_{V'V} \mathrm{d}t = \int_J \langle \partial_t v, w_m \rangle_{V'V} \mathrm{d}t,$$

so that $\lim_{m\to\infty} \lim_{n\to\infty} \int_J (\partial_t v_n, w_m)_L \, \mathrm{d}t = \int_J \langle \partial_t v, w \rangle_{V'V} \mathrm{d}t$, and we use similar arguments for the other terms. □

Exercises

Exercise 64.1 (Strong measurability). Prove the statement made in Example 64.8. (*Hint*: use Theorem 1.17.)

Exercise 64.2 (Bochner integral). Let $f : J \to V$ be a Bochner integrable function and let $(f_n)_{n\in\mathbb{N}}$ be a countable sequence of simple functions satisfying the assumptions of Definition 64.11. (i) Show that $\int_J f_n(t)\mathrm{d}t$ has a limit when $n \to \infty$. (*Hint*: prove that it is a Cauchy sequence.) (ii) Show that if $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ are two sequences of simple functions satisfying the assumptions of Definition 64.11, then $\lim_{n\to\infty}\int_J f_n(t)\mathrm{d}t = \lim_{n\to\infty}\int_J g_n(t)\mathrm{d}t$.

Exercise 64.3 ($L^p(J;V)$). Let f be a Bochner integrable function. (i) Prove that $\|\int_J f(t)\mathrm{d}t\|_V \le \int_J \|f(t)\|_V\mathrm{d}t$. (ii) Prove that $L^p(J;V) \hookrightarrow L^1(J;V)$. (iii) Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $L^1(J;V)$ s.t. $(f_n(t))_{n\in\mathbb{N}}$ converges to $f(t)$ in V and $\|f_n(t)\|_V \le g(t)$ with $g \in L^1(J;\mathbb{R})$ for a.e. $t \in J$. Show that $f \in L^1(J;V)$ and $(f_n)_{n\in\mathbb{N}}$ converges to f in $L^1(J;V)$.

Exercise 64.4 ($L^q((0,1);L^p(0,1))$). Let $p \in [1,\infty)$. Let $J := (0,1)$ and $g : J \to L^p(D)$ with $D := (0,1)$ be defined by $g(t) := \mathbb{1}_{(0,t)}$ for all $t \in J$. (i) Show that g is almost separably valued. (ii) Show that g is weakly measurable. (iii) Let $q \in [1,\infty]$. Show that $g \in L^q(J;V)$ and compute $\|g\|_{L^q(J;V)}$.

Exercise 64.5 (Constants). Let V be a Banach space and $f \in L^1_{\mathrm{loc}}(J;V)$. Assume that f is weakly differentiable and $\partial_t f = 0$. Show that there is $a \in V$ such that $f(t) = a$ a.e. $t \in J$. (*Hint*: see the proof of Lemma 2.11.)

Exercise 64.6 (Linear map). Prove Lemma 64.34.

Exercise 64.7 ($X^{p,q}(J;V,W)$). Prove that $X^{p,q}(J;V,W)$ is a Banach space.

Exercise 64.8 (Continuous embedding). Let $J := (0,T)$, $T > 0$. The goal is to prove that $X^{p,q}(J;V,W) \hookrightarrow C^0(\overline{J};W)$. Let $u \in X^{p,q}(J;V,W)$. Set $v(t) := \partial_t u(t)$ and $w(t) := \int_0^t v(\tau)\,\mathrm{d}\tau$. (i) Show that $w \in C^0(\overline{J};W)$. (*Hint*: use Lebesgue's dominated convergence theorem.) (ii) Let $\rho(\tau) := \eta e^{-\frac{1}{1-|\tau|^2}}$ if $|\tau| \le 1$ and $\rho(\tau) := 0$ otherwise, with η s.t. $\int_{\mathbb{R}}\rho(\tau)\,\mathrm{d}\tau = 1$. Let $0 < s < t < T$ and let N be the smallest integer s.t. $N \ge \max(\frac{1}{s}, \frac{1}{T-t})$. Define $\rho_n(\tau) := n\rho(n\tau)$ for all $n \ge N$. Consider the sequence of smooth functions $\phi_n(\tau) := \int_0^\tau(\rho_n(s-\xi) - \rho_n(t-\xi))\,\mathrm{d}\xi$. What is $\lim_{n\to\infty}\phi_n(\tau)$? (*Hint*: $\int_{\mathbb{R}}\rho_n(s-\xi)f(\xi)\,\mathrm{d}\xi \to f(s)$ for a.e. s and all $f \in L^1(\mathbb{R})$.) (iii) Show that $\delta_n(s,t) := \int_{-1}^1 \rho_n(y)(u(s-\frac{y}{n}) - u(t-\frac{y}{n}))\,\mathrm{d}y = -\int_0^T v(\tau)\phi_n(\tau)\,\mathrm{d}\tau$. (iv) Compute $\lim_{n\to\infty}\delta_n(s,t)$. (*Hint*: pass to the limit in the above equality and accept as a fact that $\lim_{n\to\infty}\int_{-1}^1\rho(\tau)f(s-\frac{\tau}{n})\,\mathrm{d}\tau = f(s)$ for a.e. s and all $f \in L^1(J;B)$, where B is either V or W.) (v) Prove that $u \in C^0(\overline{J};W)$ and $u \in C^{0,\frac{q-1}{q}}(\overline{J};W)$ if $q > 1$.

Exercise 64.9 (Time derivative of product). Let $\alpha \in C^1(\overline{J};\mathbb{R})$ and $u \in X^{p,q}(J;V,W)$. Show that $\partial_t(\alpha u) = u\partial_t\alpha + \alpha\partial_t u$ (see Definition 64.35).

Chapter 65

Weak formulation and well-posedness

Let D be a Lipschitz domain in $\mathbb{R}^d$, and let $J := (0,T)$ with $T > 0$ be a bounded time interval. The prototypical example of a parabolic equation is the *heat equation* which in strong form is formulated as follows:

$$\partial_t u - \nabla\cdot(\kappa\nabla u) = f \quad \text{in } D\times J, \tag{65.1a}$$
$$u_{|\partial D\times J} = 0 \quad \text{(boundary condition)}, \tag{65.1b}$$
$$u_{|D\times\{0\}} = u_0 \quad \text{(initial condition)}. \tag{65.1c}$$

The unknown is the space-time function $u : D\times J \to \mathbb{R}$, and the data are the source term $f : D\times J \to \mathbb{R}$, the initial condition $u_0 : D \to \mathbb{R}$, and the diffusion coefficient $\kappa : D\times J \to \mathbb{R}$. The goal of this chapter is to derive a weak formulation and to establish the well-posedness of a model parabolic problem in a slightly more general form than (65.1). To this purpose, we use the Bochner integration theory presented in the previous chapter.

65.1 Weak formulation

In this section, we introduce an abstract parabolic problem with the same generic properties as the heat equation and we derive a weak (and an ultraweak) formulation of this problem.

65.1.1 Heuristic argument for the heat equation

Let us assume for the time being that the solution to (65.1) is smooth. Let us proceed informally. We multiply (65.1a) by some smooth space-time function v compactly supported in $D\times J$, integrate over the space-time cylinder $D\times J$, and integrate by parts in space. We obtain

$$\int_J\int_D v\partial_t u\,\mathrm{d}x\mathrm{d}t + \int_J\int_D \kappa\nabla u\cdot\nabla v\,\mathrm{d}x\mathrm{d}t = \int_J\int_D fv\,\mathrm{d}x\mathrm{d}t, \tag{65.2}$$

which we rewrite as $\mathfrak{T}_1 + \mathfrak{T}_2 = \mathfrak{T}_3$. Since the Cauchy–Schwarz inequality implies that $|\mathfrak{T}_2| \le \|\kappa\|_{L^\infty(D\times J)}\|u\|_{L^2(J;H^1(D))}\|v\|_{L^2(J;H^1(D))}$, a natural idea to make sense of $\mathfrak{T}_2$ is to look for the

solution to (65.1) in $L^2(J;H_0^1(D))$. Moreover, by assuming that the test functions v are in $L^2(J;H_0^1(D))$, we can make sense of $\mathfrak{T}_1$ by looking for a solution such that $\partial_t u \in L^2(J;H^{-1}(D))$ and writing $\mathfrak{T}_1$ as $\int_J \langle \partial_t u(t), v(t)\rangle_{H^{-1},H_0^1}\mathrm{d}t$ since $H^{-1}(D) := (H_0^1(D))'$. Finally, $\mathfrak{T}_3$ makes sense if we assume that $f \in L^2(J;H^{-1}(D))$. In conclusion, we use the functional spaces introduced in §64.2.2 and look for $u \in X(J;H_0^1(D),H^{-1}(D))$. Then the boundary condition (65.1b) is satisfied for a.e. $t \in J$ since $u(t) \in H_0^1(D)$ for a.e. $t \in J$, and Lemma 64.40 implies that $u \in C(\overline{J};L^2(D))$, so that the initial condition $u_{|D\times\{0\}} = u_0$ makes sense provided u_0 is in $L^2(D)$.

65.1.2 Abstract parabolic problem

We now reformulate what we have done in §65.1.1 in an abstract setting that will allow us to treat a general class of equations like the heat equation. Let $V \hookrightarrow L$ be two separable real Hilbert spaces with continuous and dense embedding forming the Gelfand triple

$$V \hookrightarrow L \equiv L' \hookrightarrow V'. \tag{65.3}$$

Inspired by the functional spaces introduced in §64.2.2, we set

$$X := X(J;V,V') = \{v \in L^2(J;V) \mid \partial_t v \in L^2(J;V')\}. \tag{65.4}$$

Let $A : J \to \mathcal{L}(V;V')$ be an operator satisfying the following properties:

$$J \ni t \longmapsto \langle A(t)(v), w\rangle_{V',V} \in \mathbb{R} \text{ is measurable for all } v,w \in V, \tag{65.5a}$$

$$\exists M > 0, \quad \|A(t)(v)\|_{V'} \le M\|v\|_V, \qquad \forall v \in V, \text{ for a.e. } t \in J, \tag{65.5b}$$

$$\exists \alpha > 0, \quad \langle A(t)(v), v\rangle_{V',V} \ge \alpha\|v\|_V^2, \qquad \forall v \in V, \text{ for a.e. } t \in J. \tag{65.5c}$$

It is implicitly understood in what follows that M (resp., α) is the smallest (resp., largest) constant such that (65.5b) (resp., (65.5c)) holds true.

Lemma 65.1 (Strong measurability). *Let $p \in [1,\infty]$ and $u \in L^p(J;V)$. Let $A : J \to \mathcal{L}(V;V')$ and assume* (65.5a) *and* (65.5b). *Define the function*

$$A(u) : J \ni t \longmapsto A(u)(t) := A(t)(u(t)) \in V'. \tag{65.6}$$

Then $A(u) : J \to V'$ is strongly measurable, and $A(u) \in L^p(J;V')$ with $\|A(u)\|_{L^p(J;V')} \le M\|u\|_{L^p(J;V)}$.

Proof. We prove the strong measurability of $A(u)$ by using the Pettis measurability theorem (Theorem 64.4). $A(u)$ is almost separably valued since we assumed that V' is separable. Let us now show that for every $w \in (V')'$, the function $J \ni t \mapsto \langle w, A(u)(t)\rangle_{(V')',V'}$ is measurable. Since we have assumed that V is a Hilbert space, we can identify $(V')'$ with V, and the above property reduces to showing that $J \ni t \mapsto \langle A(u)(t), w\rangle_{V',V}$ is measurable for all $w \in V$. Since $u \in L^p(J;V)$, we infer that u is Bochner integrable, i.e., there exists a countable sequence of simple functions $(v_n)_{n\in\mathbb{N}}$ s.t. $\lim_{n\to\infty} v_n(t) = u(t)$ for a.e. $t \in J$. Since v_n is a simple function for all n, there exists a finite index set $\mathcal{I}_n \subset \mathbb{N}$, a collection of disjoint measurable subsets $\{J_{n,k}\}_{k\in\mathcal{I}_n}$ in J, and a collection of vectors $\{w_{n,k}\}_{k\in\mathcal{I}_n}$ in V s.t. $v_n(t) := \sum_{k\in\mathcal{I}_n} w_{n,k}\mathbb{1}_{J_{n,k}}(t)$ for a.e. $t \in J$. Owing to (65.5a), we infer that $J \ni t \mapsto \langle A(t)(w_{n,k}), w\rangle_{V',V}$ is measurable for all $n \in \mathbb{N}$ and all $k \in \mathcal{I}_n$. It follows that $J \ni t \mapsto \mathbb{1}_{J_{n,k}}(t)\langle A(t)(w_{n,k}), w\rangle_{V',V}$ is also measurable (because the product of two measurable functions is measurable; see Theorem 1.16). Hence, the function $J \ni t \mapsto \langle A(t)(v_n(t)), w\rangle_{V',V} = \sum_{k\in\mathcal{I}_n}\langle A(t)(w_{n,k}), w\rangle_{V',V}\mathbb{1}_{J_{n,k}}(t)$ is measurable (a finite sum of measurable functions is measurable; see Theorem 1.16). Using the boundedness property (65.5b),

we infer that $\lim_{n\to\infty}\langle A(t)(v_n(t)),w\rangle_{V',V} \to \langle A(t)(u(t)),w\rangle_{V',V}$ for a.e. $t\in J$. Invoking Theorem 1.12(ii), we deduce that $J\ni t\mapsto \langle A(t)(u(t)),w\rangle_{V',V}$ is measurable. We can now conclude that $J\ni t\mapsto A(t)(u(t))\in V'$ is strongly measurable by invoking the Pettis measurability theorem. Finally, we prove that $A(u)\in L^p(J;V')$ with $\|A(u)\|_{L^p(J;V')}\le M\|u\|_{L^p(J;V)}$ by invoking Bochner's theorem (see Exercise 65.1). □

Let $f\in L^2(J;V')$ and $u_0\in L$. The model problem we want to solve is to find $u\in X:=X(J;V,V')$ s.t.

$$\partial_t u(t)+A(u)(t)=f(t) \qquad \text{in } L^2(J;V'), \tag{65.7a}$$
$$u(0)=u_0 \qquad \text{in } L. \tag{65.7b}$$

The initial condition (65.7b) is meaningful since $X\hookrightarrow C^0(\overline{J};L)$ owing to Lemma 64.40(i). Moreover, both $\partial_t u$ and $A(u)$ are in $L^2(J;V')$ since $u\in X$.

Remark 65.2 (Real vs. complex). Working with real Hilbert spaces is natural for the heat equation. It is possible to extend the abstract theory of parabolic problems to complex spaces by replacing the assumption (65.5c) by $\Re\big(\langle A(t)(v),v\rangle_{V',V}\big)\ge\alpha\|v\|_V^2$ for all $v\in V$ and a.e. $t\in J$. □

65.1.3 Weak formulation

To reformulate (65.7) in weak form, we consider the trial space X and the test space Y such that

$$X:=X(J;V,V')=\{v\in L^2(J;V)\mid \partial_t v\in L^2(J;V')\}, \tag{65.8a}$$
$$Y:=Y_0\times Y_1, \qquad Y_0:=L, \qquad Y_1:=L^2(J;V). \tag{65.8b}$$

Notice that $L^2(J;V')\equiv L^2(J;V)'=Y_1'$ owing to Lemma 64.20(i). We define the bilinear form $b:X\times Y\to\mathbb{R}$ and the linear form $\ell:Y\to\mathbb{R}$ s.t. for all $v\in X$ and all $y:=(y_0,y_1)\in Y$,

$$b(v,y):=(v(0),y_0)_L+\int_J\langle\partial_t v(t)+A(v)(t),y_1(t)\rangle_{V',V}\mathrm{d}t, \tag{65.9a}$$
$$\ell(y):=(u_0,y_0)_L+\int_J\langle f(t),y_1(t)\rangle_{V',V}\mathrm{d}t. \tag{65.9b}$$

The definitions (65.9a) and (65.9b) are meaningful since the forms b and ℓ are bounded on $X\times Y$ and on Y, respectively. We notice that the first component $y_0\in L$ is used to enforce (65.7b) and the second component $y_1\in L^2(J;V)$ is used to enforce (65.7a). In conclusion, (65.7) is reformulated as follows:

$$\begin{cases}\text{Find } u\in X \text{ such that}\\ b(u,y)=\ell(y), \quad \forall y\in Y.\end{cases} \tag{65.10}$$

Definition 65.3 (Parabolic equation). *Let $f\in L^2(J;V')$ and $u_0\in L$. We say that the model problem* (65.10) *is* parabolic *if the operator $A:J\to\mathcal{L}(V;V')$ satisfies* (65.5).

Lemma 65.4 (Weak solution). *Let $u\in X$ solve* (65.10). *Then $\partial_t u(t)+A(u)(t)=f(t)$ in V' for a.e. $t\in J$ and $u(0)=u_0$ in L.*

Proof. Let $\phi\in C_0^\infty(J;\mathbb{R})$, let $v\in V$, and consider the test function $y:=(0,\phi v)\in Y$ in (65.10). We infer that

$$\int_J\phi(t)\langle\partial_t u(t)+A(u)(t)-f(t),v\rangle_{V',V}\mathrm{d}t=0.$$

The function $g(t) := \langle \partial_t u(t) + A(u)(t) - f(t), v\rangle_{V',V}$ is in $L^2(J;\mathbb{R})$ since $\partial_t u + A(u) - f$ is a strongly measurable function (see Lemma 65.1) such that $\|\partial_t u + A(u) - f\|_{L^2(J;V')} \le \|\partial_t u\|_{L^2(J;V')} + \|A(u)\|_{L^2(J;V')} + \|f\|_{L^2(J;V')} < \infty$. Since ϕ is arbitrary in $C_0^\infty(J;\mathbb{R})$, the vanishing integral theorem (see Theorem 1.32) implies that g vanishes a.e. in J. This proves that $\partial_t u(t) + A(u)(t) = f(t)$ in V' since the test function v is arbitrary in V. Finally, considering the test function $y := (y_0, 0) \in Y$ with y_0 arbitrary in L readily yields $u(0) = u_0$. □

Remark 65.5 (Variant with Gårding inequality). A slightly more general assumption than (65.5c) is to assume that the following *Gårding inequality* holds true: There are $\alpha > 0$ and $\eta > 0$ such that

$$\langle A(t)v, v\rangle_{V',V} \ge \alpha\|v\|_V^2 - \eta\|v\|_L^2, \qquad \forall v \in V, \text{ for a.e. } t \in J.$$

If this is the case, one can rescale the solution u and the right-hand side f of the original problem by setting $z := e^{-\eta t}u$ and $g := e^{-\eta t}f$ so as to obtain a parabolic problem for z with source term g. □

Remark 65.6 (Strong enforcement of initial condition). It is also possible to consider a functional setting where the initial condition is strongly enforced. If $u_0 = 0$, one simply considers the subspace $X_{\rm ic} := \{v \in X \mid v(0) = 0\}$, which is closed in X since $X \hookrightarrow C^0(\overline{J};L)$. The weak formulation then consists of seeking $u \in X_{\rm ic}$ s.t. $b_{\rm ic}(u,y) = \ell_{\rm ic}(y)$ for all $y \in Y_{\rm ic} := L^2(J;V)$, where $b_{\rm ic}(v,y) := b(v,(0,y))$ and $\ell_{\rm ic}(y) := \ell((0,y))$. The general case of a nonzero initial condition $u_0 \in L$ can be handled by using the surjectivity of the trace map $\gamma_0 : X \ni v \mapsto \gamma_0(v) := v(0) \in L$ (see Lemma 64.40(ii)). Letting $v_0 \in X$ be s.t. $\gamma_0(v_0) = u_0$, we then look for $u' \in X_{\rm ic}$ s.t. $b_{\rm ic}(u',y) = \ell'_{\rm ic}(y)$ for all $y \in Y_{\rm ic}$, where $\ell'_{\rm ic}(y) := \ell_{\rm ic}(y) - \int_J \langle \partial_t v_0(t) + A(v_0)(t), y(t)\rangle_{V',V}\mathrm{d}t$. Note that $\ell'_{\rm ic}$ is bounded on $Y_{\rm ic}$. Once the solution $u' \in X_{\rm ic}$ to the above problem is found, the solution to the original problem is $u := u' + v_0$. □

65.1.4 Example: the heat equation

In the context of the heat equation (65.1), the Gelfand triple is

$$V := H_0^1(D), \quad L := L^2(D) \equiv L^2(D)', \quad V' = H^{-1}(D). \tag{65.11}$$

We equip the space $H_0^1(D)$ with the H^1-seminorm, and we recall the Poincaré–Steklov inequality $C_{\rm PS}\|v\|_{L^2(D)} \le \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in H_0^1(D)$, where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. Let us assume that the diffusivity $\kappa : D\times J \to \mathbb{R}$ is continuous w.r.t. $t \in J$ and bounded over $D\times J$, and that there is $\kappa_\flat > 0$ s.t. $\kappa \ge \kappa_\flat$ a.e. in $D\times J$. The operator $A : J \to \mathcal{L}(H_0^1(D);H^{-1}(D))$ is s.t. $A(t)(v) := -\nabla\cdot(\kappa(t)\nabla v) \in H^{-1}(D)$ for all $v \in H_0^1(D)$ and a.e. $t \in J$. The assumption (65.5a) is satisfied since the function $J \ni t \mapsto \int_D \kappa(x,t)\nabla v(x)\cdot\nabla w(x)\,\mathrm{d}x$ is continuous and thus measurable for all $v, w \in H_0^1(D)$. The assumption (65.5b) is satisfied with $M := \|\kappa\|_{L^\infty(D\times J)}$. Finally, the coercivity assumption (65.5c) is satisfied with $\alpha := \kappa_\flat$.

Consider a source term $f \in L^2(J;H^{-1}(D))$ and an initial condition $u_0 \in L^2(D)$. The weak formulation of the heat equation fits the abstract form (65.10) with the functional spaces

$$X := \{v \in L^2(J;H_0^1(D)) \mid \partial_t v \in L^2(J;H^{-1}(D))\}, \tag{65.12a}$$

$$Y := L^2(D)\times L^2(J;H_0^1(D)), \tag{65.12b}$$

and the forms b, ℓ such that for all $(v,y) \in X\times Y$,

$$b(v,y) := (v(0),y_0)_{L^2(D)} + \int_J \Big(\langle \partial_t v(t), y_1(t)\rangle + (\kappa(t)\nabla v(t), \nabla y_1(t))_{\boldsymbol{L}^2(D)}\Big)\mathrm{d}t,$$

$$\ell(y) := (u_0,y_0)_{L^2(D)} + \int_J \langle f(t), y_1(t)\rangle \mathrm{d}t,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H^{-1}(D)$ and $H^1_0(D)$.

Remark 65.7 (Extensions). The theory developed in this chapter goes well beyond the heat equation since it allows one to solve the time-dependent version of all the problems studied in Chapter 31. For instance, one can consider the time-dependent differential operator $A(t)(v) := -\nabla\cdot(\mathbb{d}(\cdot,t)\nabla v) + \boldsymbol{\beta}(\cdot,t)\cdot\nabla v + \mu(\cdot,t)v$. The well-posedness results presented below still apply in this case provided the space-time fields $\mathbb{d}$, $\boldsymbol{\beta}$, and μ are continuous w.r.t. $t \in J$, bounded over $D\times J$, and s.t. $A(t)$ satisfies the coercivity property (65.5c); see Proposition 31.8 for sufficient conditions yielding coercivity. The theory generalizes to Neumann and Robin boundary conditions as well. □

65.1.5 Ultraweak formulation

It is also possible to consider a weak formulation where the smoothness of the weak time derivative is enforced on the test functions and not on the trial functions. In this setting, the trial and test spaces are

$$X_{\mathrm{uw}} := L^2(J;V), \tag{65.13a}$$

$$Y_{\mathrm{uw}} := \{w \in L^2(J;V) \mid \partial_t w \in L^2(J;V'),\ w(T) = 0\}. \tag{65.13b}$$

(Notice that $X_{\mathrm{uw}} = Y$ and $Y_{\mathrm{uw}} = \{w \in X \mid w(T) = 0\}$.) Setting

$$b_{\mathrm{uw}}(v,w) := \int_J \langle v(t), -\partial_t w(t) + A^*(w)(t)\rangle_{V,V'}\mathrm{d}t, \tag{65.14a}$$

$$\ell_{\mathrm{uw}}(w) := (u_0, w(0))_L + \int_J \langle f(t), w(t)\rangle_{V',V}\mathrm{d}t, \tag{65.14b}$$

with $A^*(w)(t) := A(t)^*(w(t))$ (the same argument as in Lemma 65.1 shows that $A^*(w) \in L^2(J;V')$), the *ultraweak formulation* is as follows:

$$\begin{cases} \text{Find } u \in X_{\mathrm{uw}} \text{ such that} \\ b_{\mathrm{uw}}(u,w) = \ell_{\mathrm{uw}}(w), \quad \forall w \in Y_{\mathrm{uw}}. \end{cases} \tag{65.15}$$

Although the ultraweak formulation (65.15) uses a larger trial space and a smaller test space than the weak formulation (65.10), the two formulations are equivalent.

Lemma 65.8 (Equivalence). (65.10) *and* (65.15) *have the same solution sets.*

Proof. (1) Assume that $u \in X$ solves (65.10). Then $u \in X_{\mathrm{uw}}$ since $X \subset X_{\mathrm{uw}}$. Moreover, since $Y_{\mathrm{uw}} \subset Y$, we have for all $w \in Y_{\mathrm{uw}}$,

$$\begin{aligned} b_{\mathrm{uw}}(u,w) &= \int_J \langle u(t), -\partial_t w(t) + A^*(w)(t)\rangle_{V,V'}\mathrm{d}t \\ &= \int_J \langle \partial_t u(t) + A(u)(t), w(t)\rangle_{V',V}\mathrm{d}t + (u(0), w(0))_L \\ &= b(u,(w(0),w)) = \ell(w(0),w) = \ell_{\mathrm{uw}}(w), \end{aligned}$$

where we used integration by parts in time (Lemma 64.40) since $u, w \in X(J;V,V')$ and $w(T) = 0$ since $w \in Y_{\mathrm{uw}}$. (Note that $\langle u(t), A^*(w)(t)\rangle_{V,V'} = \langle u(t), A(t)^*(w(t))\rangle_{V,V'} = \langle A(t)(u(t)), w(t)\rangle_{V',V} = \langle A(u)(t), w(t)\rangle_{V',V}$.)

(2) Assume that $u \in X_{\rm uw}$ solves (65.15). Let $\phi \in C_0^\infty(J;\mathbb{R})$ and $v \in V$. Notice that $\phi v \in Y_{\rm uw}$, so that $b_{\rm uw}(u,\phi v) = \ell_{\rm uw}(\phi v)$. This yields

$$\int_J -(u(t),v)_L \phi'(t)\mathrm{d}t = \int_J \langle f(t) - A(u)(t), v\rangle_{V',V}\phi(t)\mathrm{d}t.$$

Owing to Proposition 64.33, this identity shows that u has a weak time derivative in $L^2(J;V')$, i.e., u is a member of X, and it also shows that $\partial_t u = f - A(u)$ in $L^2(J;V')$. Let now $\phi \in C^\infty(\overline{J};\mathbb{R})$ with $\phi(T) = 0$, so that ϕv is again in $Y_{\rm uw}$. Integrating by parts in time (Lemma 64.40) and using $\partial_t u = f - A(u)$, the identity $b_{\rm uw}(u,\phi v) = \ell_{\rm uw}(\phi v)$ yields $\phi(0)(u(0),v)_L = \phi(0)(u_0,v)_L$. Choosing ϕ s.t. $\phi(0) = 1$, and since v is arbitrary in V which is dense in L, we infer that $u(0) = u_0$. □

65.2 Well-posedness

The objective of this section is to establish the well-posedness of the parabolic model problem (65.10). More precisely, we prove the following result.

Theorem 65.9 (Lions). *The problem* (65.10) *is well-posed under the assumption* (65.5).

This result has been established in Lions [218, Thm. 2.1, p. 219]; see also Lions and Magenes [220, Thm. 4.1, p. 238] or Dautray and Lions [100, Thm. 2, p. 513]. We prove this result in two steps. We first establish the uniqueness of the solution using a coercivity-like argument. Then we use a constructive argument to establish the existence of the weak solution. In Chapter 71, we revisit the whole well-posedness argument in the context of the BNB theorem (Theorem 25.9) by establishing an inf-sup condition.

65.2.1 Uniqueness using a coercivity-like argument

In this section, we show that (65.10) admits at most one solution $u \in X$. This is done by establishing an a priori estimate on the weak solution, that is, by showing that the weak solution depends continuously on the data f and u_0. The continuous dependence is established by invoking a coercivity-like argument where we use $(0,u) \in Y$ as the test function in (65.10).

Lemma 65.10 (A priori estimate and uniqueness). *Assume that the function $u \in X$ solves the parabolic problem* (65.10). (i) *The following a priori estimate holds true:*

$$\alpha\|u\|^2_{L^2(J;V)} + \|u(T)\|^2_L \le \frac{1}{\alpha}\|f\|^2_{L^2(J;V')} + \|u_0\|^2_L. \tag{65.16}$$

(ii) *The model problem* (65.10) *admits at most one solution.*

Proof. (1) Proof of (65.16). Owing to the time integration by parts formula from Lemma 64.40, we infer that $\int_J \langle \partial_t u(t), u(t)\rangle_{V',V}\mathrm{d}t = \frac{1}{2}\|u(T)\|^2_L - \frac{1}{2}\|u_0\|^2_L$, where we used that $u(0) = u_0$. Moreover, the coercivity property (65.5c) implies that $\alpha\|u\|^2_{L^2(J;V)} \le \int_J \langle A(u)(t), u(t)\rangle_{V',V}\mathrm{d}t$ (recall that $A(u)(t) = A(t)(u(t))$). Putting these two identities together, we infer that

$$\begin{aligned}\alpha\|u\|^2_{L^2(J;V)} + \frac{1}{2}\|u(T)\|^2_L - \frac{1}{2}\|u_0\|^2_L &\le \int_J \langle \partial_t u(t) + A(u)(t), u(t)\rangle_{V',V}\mathrm{d}t\\ &= b(u,(0,u)) = \ell((0,u)) = \int_J \langle f(t), u(t)\rangle_{V',V}\mathrm{d}t\\ &\le \|f\|_{L^2(J;V')}\|u\|_{L^2(J;V)} \le \frac{\alpha}{2}\|u\|^2_{L^2(J;V)} + \frac{1}{2\alpha}\|f\|^2_{L^2(J;V')},\end{aligned}$$

where we used Young's inequality in the last bound. Rearranging the terms leads to (65.16).
(2) Proof of uniqueness. Assume that u_1, u_2 are two solutions to (65.10). By linearity, the difference $\delta := u_1 - u_2$ solves the parabolic problem (65.10) with data $f = 0$ and $u_0 = 0$. The a priori estimate (65.16) implies that $\delta = 0$, i.e., $u_1 = u_2$. Therefore, (65.10) admits at most one solution. □

The estimate (65.16) implies that $\|u(T)\|_L^2 \le \frac{1}{\alpha}\|f\|_{L^2(J;V')}^2 + \|u_0\|_L^2$. This estimate is not very sharp since it does not capture the important property of parabolic problems that the influence of the initial condition decays exponentially fast in time.

Lemma 65.11 (L-norm estimate, exponential decay). *Assume that u solves* (65.10)*. The following holds true for all $t \in (0,T]$ with $J_t := (0,t)$:*

$$\|u(t)\|_L^2 \le \frac{1}{\alpha}\|e^{-\frac{t-\cdot}{\rho}} f\|_{L^2(J_t;V')}^2 + e^{-2\frac{t}{\rho}}\|u_0\|_L^2, \tag{65.17}$$

with the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, i.e., the smallest constant s.t. $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$.

Proof. See Exercise 65.4. □

Example 65.12 (Heat equation). In the context of the heat equation (see §65.1.4) where $\alpha := \kappa_\flat$, Lemma 65.10 yields

$$\kappa_\flat\|u\|_{L^2(J;H_0^1(D))}^2 + \|u(T)\|_{L^2(D)}^2 \le \frac{1}{\kappa_\flat}\|f\|_{L^2(J;H^{-1}(D))}^2 + \|u_0\|_{L^2(D)}^2,$$

and defining the time scale $\rho := \frac{2}{C_{\mathrm{ps}}^2}\frac{\ell_D^2}{\kappa_\flat}$ ($\iota_{L,V} = \frac{\ell_D}{C_{\mathrm{ps}}}$ owing to the Poincaré–Steklov inequality), Lemma 65.11 yields for all $t \in (0,T]$ with $J_t := (0,t)$,

$$\|u(t)\|_{L^2(D)}^2 \le \frac{1}{\kappa_\flat}\|e^{-\frac{t-\cdot}{\rho}} f\|_{L^2(J_t;H^{-1}(D))}^2 + e^{-2\frac{t}{\rho}}\|u_0\|_{L^2(D)}^2.$$ □

65.2.2 Existence using a constructive argument

The existence of a solution to the problem (65.10) is done by invoking a semi-discrete Galerkin-type argument.

Lemma 65.13 (Existence). *There exists $u \in X$ solving the parabolic problem* (65.10)*.*

Proof. Let $(v_i)_{i\in\mathbb{N}}$ be a Hilbert basis of V (see Definition 46.19 and Theorem 46.21). Let us set $V_n := \mathrm{span}\{v_i\}_{i\in\{0:n\}}$ for all $n \in \mathbb{N}$ and let $\Pi_n : L \to V_n$ be the orthogonal projection onto V_n in L. With $u_n(0) := \Pi_n(u_0)$, consider now the following finite set of coupled linear ordinary differential equations:

$$(\partial_t u_n(t), v_i)_L + \langle A(t)(u_n(t)), v_i\rangle_{V',V} = \langle f(t), v_i\rangle_{V',V}, \quad \forall i \in \{0{:}n\}. \tag{65.18}$$

Let us set $u_n(t) := \sum_{j\in\{0:n\}} \mathsf{U}_{jn}(t)v_j$, $\mathcal{A}_{ij}(t) := \langle A(t)(v_j), v_i\rangle_{V',V}$, $\mathsf{F}_i(t) := \langle f(t), v_i\rangle_{V',V}$, $\mathcal{M}_{ij} := (v_j, v_i)_L$ for all $i,j \in \{0{:}n\}$. Defining the $\mathbb{R}^{n+1}$-valued vectors $\mathsf{U}_n(t) := (\mathsf{U}_{0n},\dots,\mathsf{U}_{nn}(t))^\mathsf{T}$, $\mathsf{F}_n(t) := (\mathsf{F}_0,\dots,\mathsf{F}_n(t))^\mathsf{T}$, the above system is equivalent to

$$\mathcal{M}\partial_t\mathsf{U}_n(t) + \mathcal{A}(t)\mathsf{U}_n(t) = \mathsf{F}_n(t), \quad \text{for a.e. } t \in J.$$

Owing to the boundedness assumption (65.5b), the Cauchy–Lipschitz theorem implies that the above problem has a unique solution (see, e.g., Brezis [52, Thm. 7.3]). Multiplying (65.18) by $\mathsf{U}_{in}(t)$, summing over $i \in \{0{:}n\}$, and integrating over J, we infer that (see the proof of Lemma 65.10)

$$\alpha\|u_n\|_{L^2(J;V)}^2 + \|u_n(T)\|_L^2 \le \frac{1}{\alpha}\|f\|_{L^2(J;V')}^2 + \|u_0\|_L^2.$$

This shows that the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $L^2(J;V)$. Since $L^2(J;V)$ is a Hilbert space (hence reflexive), Theorem C.23 implies the existence of a subsequence, which we abusively denote again by $(u_n)_{n\in\mathbb{N}}$, that converges weakly to some $u \in L^2(J;V)$, i.e., $u_n \rightharpoonup u$ in $L^2(J;V)$ (and hence also in $L^2(J;L)$).

It remains to show that u solves (65.10). Let $\phi \in C^\infty(\overline{J};\mathbb{R})$ with $\phi(T) = 0$. We multiply (65.18) by $\phi(t)$, integrate over J, integrate by parts in time, and use the linearity of $A(t)^*$ and $u_n(0) = \Pi_n(u_0)$ to obtain

$$\begin{aligned}\int_J \langle f(t), \phi(t)v_i\rangle_{V',V}\mathrm{d}t &= \int_J \Big((\partial_t u_n(t), \phi(t)v_i)_L + \langle A(t)(u_n(t)), \phi(t)v_i\rangle_{V',V}\Big)\mathrm{d}t\\ &= \int_J \Big(- (u_n(t), \phi'(t)v_i)_L + \langle u_n(t), \phi(t)A(t)^*(v_i)\rangle_{V,V'}\Big)\mathrm{d}t - (u_0, \phi(0)v_i)_L.\end{aligned}$$

We can now pass to the limit $n \to \infty$ since $u_n \rightharpoonup u$ in $L^2(J;L)$ and $\phi'(t)v_i \in L^2(J;L)$, and $u_n \rightharpoonup u$ in $L^2(J;V)$ and $\phi(t)A(t)^*(v_i) \in L^2(J;V')$. Hence,

$$\int_J \langle f(t), \phi(t)v_i\rangle_{V',V}\mathrm{d}t + (u_0, \phi(0)v_i)_L = \int_J \Big((u(t), -\phi'(t)v_i)_L + \langle u(t), \phi(t)A(t)^*(v_i)\rangle_{V,V'}\Big)\mathrm{d}t.$$

Since the above equality is satisfied for all $i \in \mathbb{N}$, it is satisfied by replacing v_i by any $v \in V$. Recalling the ultraweak formulation (65.15) shows that $b_{\mathrm{uw}}(u, \phi v) = \ell_{\mathrm{uw}}(\phi v)$ for all $\phi \in C^\infty(\overline{J};\mathbb{R})$ with $\phi(T) = 0$ and all $v \in V$. Repeating the arguments from the proof of Lemma 65.8, we conclude that $u \in X$ solves (65.10). □

65.3 Maximum principle for the heat equation

Another important property of parabolic problems is the maximum principle. For simplicity, we focus on the heat equation.

Theorem 65.14 (Maximum principle, heat equation). *Let u solve the heat equation with $f \in L^2(J;L^2(D))$ and $u_0 \in L^2(D)$.*

(i) *Assume that* $\operatorname{ess\,inf}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x})) > -\infty$ *and* $f \ge 0$ *a.e. in* $D\times J$. *Then*

$$u \ge \min(0, \operatorname*{ess\,inf}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x}))) \quad \text{a.e. in } D\times J.$$

(ii) *Assume that* $\operatorname{ess\,sup}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x})) < \infty$ *and* $f \le 0$ *a.e. in* $D\times J$. *Then*

$$u \le \max(0, \operatorname*{ess\,sup}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x}))) \quad \text{a.e. in } D\times J.$$

Proof. We are going to use a technique known in the literature as *Stampacchia's truncation* method (see Brezis [52, Thm. 10.3, p. 333]). Let $G \in C^1(\mathbb{R};\mathbb{R})$ be s.t. $G_{|(-\infty,0]} := 0$, $G'_{|(0,1)} \in (0,2)$,

$G'_{|[1,\infty)} := 2$. Let $K \in C^2(\mathbb{R};\mathbb{R})$ be defined by $K(v) := \int_0^v G(\xi)\,\mathrm{d}\xi$. For instance, we can take

$$G(v) := \begin{cases} 0 & \text{if } v \le 0, \\ v^2 & \text{if } 0 \le v \le 1, \\ 2v-1 & \text{if } 1 \le v, \end{cases} \qquad K(v) := \begin{cases} 0 & \text{if } v \le 0, \\ \frac{1}{3}v^3 & \text{if } 0 \le v \le 1, \\ v^2 - v + \frac{1}{3} & \text{if } 1 \le v. \end{cases}$$

Proof of (i). Let $C := \min(0, \operatorname{ess\,inf}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x})))$. Owing to Lemma 65.15 below, we have $G(C-u) \in L^2(J;H_0^1(D))$. Let $t \in J$ and $J_t := (0,t)$. Let ℓ_t and b_t be the restrictions of ℓ and b to J_t. Since $-G(C-u)$ is an admissible test function in $L^2(J_t;H_0^1(D))$, $f \ge 0$, and $K(C-u_0) = 0$, we have

$$\begin{aligned} 0 \ge & -\int_{J_t} (f(s), G(C-u(s)))_{L^2(D)}\,\mathrm{d}s = -b_t(u,(0,G(C-u))) \\ = & -\int_{J_t} \Big(\langle \partial_s u(s), G(C-u(s))\rangle + \big(\kappa(s)\nabla u(s), \nabla G(C-u(s))\big)_{\boldsymbol{L}^2(D)} \Big)\,\mathrm{d}s. \end{aligned}$$

Using the identities (see Lemma 65.15) $\langle \partial_t u(t), G(u(t))\rangle = \partial_t \|K(u(t))\|_{L^1(D)}$ for a.e. $t \in J$ and $-\nabla u(s)\cdot\nabla G(C-u(s)) = G'(C-u(s))\|\nabla u(s)\|_{\ell^2}^2 \ge 0$, we infer that

$$\begin{aligned} 0 \ge & \|K(C-u(t))\|_{L^1(D)} - \|K(C-u_0)\|_{L^1(D)} \\ & + \int_{J_t} \big(\kappa(s)G'(C-u(s))\nabla u(s), \nabla u(s)\big)_{\boldsymbol{L}^2(D)}\,\mathrm{d}s \ge \|K(C-u(t))\|_{L^1(D)}. \end{aligned}$$

This implies that $K(C-u(t)) = 0$ since K takes nonnegative values. Hence, with an abuse of notation, we have $u(t) \ge C$. Since t is arbitrary in J, we infer that $u(\boldsymbol{x},t) \ge C$ for a.e. $(\boldsymbol{x},t)$ in $D\times J$.

Proof of (ii). We proceed as above with $C := \max(0, \operatorname{ess\,sup}_{\boldsymbol{x}\in D}(u_0(\boldsymbol{x})))$, but this time we use the test function $G(u-C)$. Notice that $G(u-C)$ is indeed a member of $L^2(J;H_0^1(D))$ owing to Lemma 65.15. □

Lemma 65.15 (Regularity of truncated functions). *Let the functions G, K be as above. Let $u \in X(J,H_0^1(D),H^{-1}(D))$.* (i) *Let $C \le 0$. Then $G(C-u) \in L^2(J,H_0^1(D))$ and $\nabla G(C-u(t)) = -G'(C-u(t))\nabla u(t)$ for a.e. $t \in J$.* (ii) *Let $C \ge 0$. Then $G(u-C) \in L^2(J,H_0^1(D))$ and $\nabla G(u(t)-C) = G'(u(t)-C)\nabla u(t)$ for a.e. $t \in J$.* (iii) *$K(u) \in W^{1,1}(J;L^1(D))$ and $\langle \partial_t u(t), G(u(t))\rangle = \partial_t \|K(u(t))\|_{L^1(D)}$ for a.e. $t \in J$.*

Proof. (i) Since $u \in L^2(J,H_0^1(D))$, we have $u(t) \in H_0^1(D)$ and $C-u \in H^1(D)$ for a.e. $t \in J$ (recall that D is bounded). Owing to Corollary 2.24, $G(C-u(t))$ is in $H^1(D)$, and since $G(C) = 0$ (recall that $C \le 0$), $G(C-u(t))$ is actually in $H_0^1(D)$. Corollary 2.24 also implies that $\nabla G(C-u(t)) = -G'(C-u(t))\nabla u(t)$ for a.e. $t \in J$. This in turn implies that $\|\nabla G(C-u(t))\|_{\boldsymbol{L}^2(D)} \le 2\|\nabla u(t)\|_{\boldsymbol{L}^2(D)}$ since $\|G'\|_{L^\infty(\mathbb{R})} \le 2$. Hence, $G(C-u) \in L^2(J;H_0^1(D))$.

(ii) The proof of the second statement is identical except that we use $G(-C) = 0$ since $C \ge 0$.

(iii) Owing to Theorem 64.36, there is a sequence $(u_n)_{n\in\mathbb{N}}$ in $C^\infty(\overline{J};V)$ s.t. $u_n \to u$ in X, i.e., $u_n \to u$ in $L^2(J;H_0^1(D))$ and $\partial_t u_n \to \partial_t u$ in $L^2(J;H^{-1}(D))$. Let $\phi \in C_0^\infty(J)$. Using that $G(s) = K'(s)$, we have

$$\int_J (G(u_n(t)), \partial_t u_n(t))_{L^2(D)}\phi(t)\mathrm{d}t = \int_J \int_D -K(u_n(t))\partial_t\phi(t)\,\mathrm{d}x\mathrm{d}t.$$

Let us now pass to the limit. Since $|G(u_n(t)) - G(u(t))| \le 2|u_n(t)-u(t)|$, we have $G(u_n) \to G(u)$ in $L^2(J;L^2(D))$. Moreover, we have $\nabla(G(u_n(t)) = G'(u_n(t))\nabla u_n(t)$. But $G'(u_n(t)) \to G'(u(t))$

and $\nabla u_n(t) \to \nabla u(t)$ for a.e. $t \in J$, and for n large enough, we also have $\|G'(u_n(t))\nabla u_n(t)\|_{\boldsymbol{L}^2(D)} \le 4\|\nabla u(t)\|_{\boldsymbol{L}^2(D)}$ (here we used that u_n converges strongly to u in $L^2(J;H_0^1(D))$). Lebesgue's dominated convergence theorem implies that

$$\nabla(G(u_n(t)) = G'(u_n(t))\nabla u_n(t) \to G'(u(t))\nabla u(t) = \nabla G(u(t)).$$

This argument proves that $G(u_n) \to G(u)$ in $L^2(J;H_0^1(D))$. Finally, the inequality $|K(u_n(t)) - K(u(t))| \le 2(|u_n(t)| + |u(t)|)|u_n(t) - u(t)|$ shows that $K(u_n) \to K(u)$ in $L^1(J;L^1(D))$. Hence, we can pass to the limit and obtain

$$\int_J \langle \partial_t u(t), G(u(t))\rangle_{H^{-1},H_0^1}\phi(t)\mathrm{d}t = -\int_J \left(\int_D K(u(t))\,\mathrm{d}x\right)\partial_t\phi(t)\mathrm{d}t.$$

Since $\phi \in C_0^\infty(J)$ is arbitrary and $K \ge 0$, we have $\langle \partial_t u(t), G(u(t))\rangle_{H^{-1},H_0^1} = \partial_t\|K(u(t))\|_{L^1(D)}$. This also proves that $K(u) \in W^{1,1}(J;L^1(D))$. □

Exercises

Exercise 65.1 (L^p-integrability of $A(u)$). Let $u \in L^p(J;V)$ and let $A(u)$ be defined in (65.6). Prove that $A(u) \in L^p(J;V')$ with $\|A(u)\|_{L^p(J;V')} \le M\|u\|_{L^p(J;V)}$. (*Hint*: use Theorem 64.12.)

Exercise 65.2 (Ultraweak formulation). Write the ultraweak formulation for the heat equation.

Exercise 65.3 (Gronwall's lemma). Let $J := (0,T)$, $T > 0$. Let $\alpha, \beta, u \in L^1(J;\mathbb{R})$ be s.t. $\alpha\beta, \beta u \in L^1(J;\mathbb{R})$, $\beta(t) \ge 0$, and $u(t) \le \alpha(t) + \int_0^t \beta(r)u(r)\,\mathrm{d}r$ for a.e. $t \in J$. (i) Prove that $v(t) := e^{-\int_0^t \beta(r)\,\mathrm{d}r}\int_0^t \beta(r)u(r)\,\mathrm{d}r$ is in $W^{1,1}(J;\mathbb{R})$. (ii) Prove that $v(t) \le \int_0^t \alpha(r)\beta(r)e^{-\int_0^r \beta(s)\,\mathrm{d}s}\,\mathrm{d}r$. (iii) Prove that

$$u(t) \le \alpha(t) + \int_0^t \alpha(s)\beta(s)e^{\int_s^t \beta(r)\,\mathrm{d}r}\,\mathrm{d}s. \tag{65.19}$$

(*Hint*: use Step (ii) and $\int_0^t \beta(r)u(r)\,\mathrm{d}r = v(t)e^{\int_0^t \beta(r)\,\mathrm{d}r}$.) (iv) Assume now that α is nondecreasing, i.e., $\alpha(r) \le \alpha(t)$ for a.e. $r,t \in J$ s.t. $r \le t$. Prove that for a.e. $t \in J$,

$$u(t) \le \alpha(t)e^{\int_0^t \beta(r)\,\mathrm{d}r}. \tag{65.20}$$

(v) Assume that β is constant and $\alpha \in W^{1,1}(J)$. Prove that for a.e. $t \in J$, $u(t) \le \alpha(0)e^{\beta t} + \int_0^t \alpha'(r)e^{\beta(t-r)}\,\mathrm{d}r$. *Note*: owing to the assumption $\beta(t) \ge 0$, Gronwall's lemma can be used to show that the function u has at most exponential growth in time, but it cannot be used to show that u has exponential decay. However, if the assumption $u(t) \le \alpha(t) + \int_0^t \beta(r)u(r)\,\mathrm{d}r$ is replaced by the stronger assumption $u'(t) \le \alpha'(t) + \beta(t)u(t)$, then $u(t) \le e^{\int_0^t \beta(r)\,\mathrm{d}r}u(0) + \int_0^t \alpha'(r)e^{\int_r^t \beta(s)\,\mathrm{d}s}\,\mathrm{d}r$ regardless of the sign of β.

Exercise 65.4 (Exponentially decaying estimate). (i) Prove the a priori estimate (65.17). (*Hint*: adapt the proof of Lemma 65.10 by considering the test function $(0,w) \in Y$ with $w(t) := e^{2\frac{t}{\rho}}u(t)$ and the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$.) (ii) Assuming that $f \in L^\infty((0,\infty);V')$, prove that $\limsup_{t\to\infty}\|u(t)\|_L \le \frac{\iota_{L,V}}{\alpha}\|f\|_{L^\infty((0,\infty);V')}$. (*Hint*: use (65.17).)

Chapter 66

Semi-discretization in space

We are concerned in this chapter with the semi-discretization in space of the model parabolic problem (65.10), that is, the approximation is done with respect to the space variable but the time variable is kept continuous. We use V-conforming finite elements for the space approximation. Error estimates are derived by invoking coercivity-like arguments. Semi-discretization in space leads to a (large) system of coupled ordinary differential equations (ODEs). This system of ODEs can then be discretized in time by many time-stepping techniques, as exemplified in the following chapters. This approach is often called *method of lines* in the literature.

66.1 Model problem

Let us briefly recall from §65.1 the setting for the model parabolic problem (65.10). We consider the Gelfand triple $V \hookrightarrow L \equiv L' \hookrightarrow V'$, the time interval $J := (0,T)$ with $T > 0$, and the functional spaces

$$X := \{v \in L^2(J;V) \mid \partial_t v \in L^2(J;V')\}, \tag{66.1a}$$

$$Y := L \times L^2(J;V). \tag{66.1b}$$

Let $f \in L^2(J;V')$ and $u_0 \in L$. Assume that the operator $A : J \to \mathcal{L}(V;V')$ satisfies the properties (65.5). In the context of finite elements, one usually works with bilinear forms. Thus, we set $a(t;v,w) := \langle A(t)(v), w\rangle_{V',V}$ for all $v, w \in V$ and a.e. $t \in J$. We consider the bilinear and linear forms

$$b(v,y) := (v(0), y_0)_L + \int_J \Big(\langle \partial_t v(t), y_1(t)\rangle_{V',V} + a(t; v(t), y_1(t))\Big)\mathrm{d}t, \tag{66.2a}$$

$$\ell(y) := (u_0, y_0)_L + \int_J \langle f(t), y_1(t)\rangle_{V',V}\mathrm{d}t, \tag{66.2b}$$

for all $v \in X$ and all $y := (y_0, y_1) \in Y$. The weak formulation of (65.10) is as follows:

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ b(u,y) = \ell(y), \quad \forall y \in Y. \end{cases} \tag{66.3}$$

Example 66.1 (Heat equation). The Gelfand triple is realized by taking $V := H_0^1(D)$, $L := L^2(D) \equiv L^2(D)'$, and $V' = H^{-1}(D)$. The space V is equipped with the H^1-seminorm, i.e.,

$\|v\|_V := |v|_{H^1(D)} = \|\nabla v\|_{\boldsymbol{L}^2(D)}$. This is legitimate owing to the Poincaré–Steklov inequality $C_{\rm PS}\|v\|_{L^2(D)} \le \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in H^1_0(D)$, where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. The bilinear form is $a(t;v,w) := \int_D \kappa(x,t)\nabla v(x)\cdot\nabla w(x)\,\mathrm{d}x$, where $\kappa : D\times J \to \mathbb{R}$ is continuous w.r.t. $t \in J$, uniformly bounded from above, and bounded from below away from zero on $D\times J$ (see §65.1.4). □

66.2 Principle and algebraic realization

In order to realize the approximation in space while keeping the time variable continuous, we introduce a sequence $(V_h)_{h\in\mathcal{H}}$ of finite-dimensional subspaces of V built using a finite element and a shape-regular mesh family $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly. In the case of the heat equation, V_h can be one of the H^1-conforming finite element spaces (see Chapter 19). Note that the mesh $\mathcal{T}_h$ used to build V_h is kept fixed in time. We assume to have at hand a basis of V_h, say $\{\varphi_i\}_{i\in\{1:I\}}$ (for instance the global shape functions). We consider the following semi-discrete trial and test spaces:

$$X_h := H^1(J;V_h) = X(J;V_h,V_h), \qquad Y_h := V_h\times L^2(J;V_h). \tag{66.4}$$

The spaces X_h and Y_h are still infinite-dimensional because the time variable is kept continuous. A generic function $v_h \in X_h$ is of the form $v_h(\boldsymbol{x},t) := \sum_{i\in\{1:I\}} \mathsf{V}_i(t)\varphi_i(\boldsymbol{x})$ with $\mathsf{V}_i \in H^1(J)$ for all $i \in \{1:I\}$. Similarly, a generic function $y_h \in Y_h$ is a pair $y_h := (y_{0h}, y_{1h})$ with $y_{0h} \in V_h$ and $y_{1h}(\boldsymbol{x},t) := \sum_{i\in\{1:I\}} \mathsf{Y}_i(t)\varphi_i(\boldsymbol{x})$ with $\mathsf{Y}_i \in L^2(J)$ for all $i \in \{1:I\}$. We observe that

$$X_h \subset X, \qquad Y_h \subset Y, \tag{66.5}$$

and in particular we have $\partial_t v_h(\boldsymbol{x},t) = \sum_{i\in\{1:I\}} \mathsf{V}_i'(t)\varphi_i(\boldsymbol{x})$. The semi-discrete counterpart of (66.3) is as follows:

$$\begin{cases} \text{Find } u_h \in X_h \text{ such that} \\ b(u_h, y_h) = \ell(y_h), \quad \forall y_h \in Y_h. \end{cases} \tag{66.6}$$

Owing to (66.5), the approximation setting is *conforming*. Since the duality pairing between V' and V is an extension of the inner product in L and since $V_h \subset V \hookrightarrow L$, we infer that the bilinear form b restricted to $X_h\times Y_h$ is s.t.

$$b(v_h, y_h) = (v_h(0), y_{0h})_L + \int_J \Big((\partial_t v_h(t), y_{1h}(t))_L + a(t; v_h(t), y_{1h}(t)) \Big)\mathrm{d}t.$$

Let $\mathcal{P}_{V_h} : L \to V_h$ be the L-orthogonal projection, i.e., for all $z \in L$, $\mathcal{P}_{V_h}(z)$ is the unique element in V_h s.t. $(z - \mathcal{P}_{V_h}(z), w_h)_L := 0$ for all $w_h \in V_h$.

Proposition 66.2 (Equivalence and well-posedness). (i) *A function $u_h \in X_h$ solves* (66.6) *iff for all $w_h \in V_h$,*

$$(\partial_t u_h(t), w_h)_L + a(t; u_h(t), w_h) = \langle f(t), w_h\rangle_{V',V} \quad \text{in } L^2(J), \tag{66.7a}$$

$$u_h(0) = \mathcal{P}_{V_h}(u_0). \tag{66.7b}$$

(ii) *The semi-discrete problems* (66.6) *and* (66.7) *are well-posed. Moreover, if $f \in C^0(\overline{J};V')$ and $A \in C^0(\overline{J};\mathcal{L}(V;V'))$, we have $u_h \in C^1(\overline{J};V_h)$.*

Proof. (i) The equivalence of (66.6) with (66.7) follows by taking first the test function $(y_{0h}, 0)$ with y_{0h} arbitrary in V_h, and then taking the test function $(0, y_{1h})$ with y_{1h} arbitrary in $L^2(J; V_h)$. (ii) Let $u_h(\boldsymbol{x}, t) := \sum_{i\in\{1:I\}} \mathsf{U}_i(t)\varphi_i(\boldsymbol{x})$ be the expansion of the semi-discrete solution $u_h \in X_h$ in the basis $\{\varphi_i\}_{i\in\{1:I\}}$. We set $\mathsf{U}(t) := (\mathsf{U}_1(t), \dots, \mathsf{U}_I(t))^\mathsf{T} \in \mathbb{R}^I$ and introduce the (time-dependent) stiffness matrix $\mathcal{A}(t) \in \mathbb{R}^{I\times I}$ and the (time-independent) mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$ such that

$$\mathcal{A}_{ij}(t) = a(t; \varphi_j, \varphi_i), \qquad \mathcal{M}_{ij} = (\varphi_j, \varphi_i)_L, \qquad \forall i, j \in \{1:I\}.$$

The mass matrix is symmetric positive definite, and the stiffness matrix is positive definite for a.e. $t \in J$; see §28.2.2-§28.2.3. Using the above notation, (66.7) is recast as follows:

$$\begin{cases} \mathcal{M}\partial_t \mathsf{U}(t) = -\mathcal{A}(t)\mathsf{U}(t) + \mathsf{F}(t) \quad \text{in } L^2(J), \\ \qquad \mathsf{U}(0) = \mathsf{U}_0, \end{cases} \tag{66.8}$$

where $\mathsf{F}(t) := (\langle f(t), \varphi_1\rangle_{V',V}, \dots, \langle f(t), \varphi_I\rangle_{V',V})^\mathsf{T} \in \mathbb{R}^I$ and $\mathsf{U}_0 \in \mathbb{R}^I$ is the coordinate vector of $\mathcal{P}_{V_h}(u_0)$ relative to the basis $\{\varphi_i\}_{i\in\{1:I\}}$. Since (66.8) is a finite coupled system of linear ODEs, the Cauchy–Lipschitz theorem guarantees the existence and uniqueness of a solution $\mathsf{U}(t)$ in $H^1(J; \mathbb{R}^I)$; see, e.g., Brezis [52, Thm. 7.3]. Finally, if $f \in C^0(\overline{J}; V')$ and $A \in C^0(\overline{J}; \mathcal{L}(V; V'))$, then (66.8) is satisfied for all $t \in \overline{J}$, and we have $\mathsf{U} \in C^1(\overline{J}; \mathbb{R}^I)$, i.e., $u_h \in C^1(\overline{J}; V_h)$. □

Example 66.3 (Duhamel's formula). If the operator A is time-independent, then so is the matrix $\mathcal{A}$, and the unique solution to (66.8) is given by $\mathsf{U}(t) = \mathsf{U}_0 + \int_0^t e^{(s-t)\mathcal{M}^{-1}\mathcal{A}}\mathcal{M}^{-1}\mathsf{F}(s)\,\mathrm{d}s$ for all $t \in J$. This expression is often called *Duhamel's formula*. □

Remark 66.4 (Initialization). Other initializations than $u_h(0) = \mathcal{P}_{V_h}(u_0)$ can be realized if one replaces b and ℓ by some other consistent approximations, say b_h and ℓ_h. For instance, leaving b unchanged, one can consider $\ell_h(y_h) := (\mathcal{I}_h(u_0), y_{0h})_L + \int_J \langle f(t), y_{1h}(t)\rangle_{V',V}\mathrm{d}t$, where $\mathcal{I}_h$ is some L-stable approximation operator. This gives $u_h(0) = \mathcal{I}_h(u_0)$. □

Remark 66.5 (Mass lumping). It is sometimes possible to replace $\mathcal{M}$ by a diagonal matrix. One possibility consists of using a quadrature (see Chapter 30) to evaluate the term involving the time derivative in (66.7a). Assume for instance that $\{\varphi_i\}_{i\in\{1:I\}}$ is a Lagrange basis associated with the nodes $\{\boldsymbol{a}_i\}_{i\in\{1:I\}}$. Then the quadrature $\int_D v_h(\boldsymbol{x})\,\mathrm{d}x = \sum_{i\in\{1:I\}} m_i v_h(\boldsymbol{a}_i)$, with $m_i := \int_D \varphi_i\,\mathrm{d}x$, is exact for all $v_h \in V_h$. Using this quadrature, one approximates the (consistent) mass matrix $\mathcal{M}$ by the diagonal matrix $\overline{\mathcal{M}}$ with diagonal entries $\{m_i\}_{i\in\{1:I\}}$. This process is called *mass lumping* and $\overline{\mathcal{M}}$ is called *lumped mass matrix*. We refer the reader to Thomée [273, Chap. 15] for the analysis of the lumping technique for parabolic problems. An equivalent viewpoint leading to the same lumped mass matrix is to consider a piecewise constant reconstruction operator from the degrees of freedom to evaluate the term with the time derivative in (66.7a); see Raviart [242]. □

Remark 66.6 (Tensor products). Using tensor-product notation (see Remark 64.24), definitions equivalent to (66.4) are $X_h := H^1(J) \otimes V_h$ and $Y_h := V_h \times (L^2(J) \otimes V_h)$. These choices are reasonable since $L^2(J) \otimes V$ is dense in $L^2(J; V)$ and $(V_h)_{h\in\mathcal{H}}$ has approximation properties in V. □

66.3 Error analysis

In this section, we perform the error analysis of the semi-discrete problem (66.6) using coercivity arguments. We bound the error in the $L^2(J; V)$-norm and in the $C^0(\overline{J}; L)$-norm, and we illustrate the error estimates in the case of the heat equation.

66.3.1 Error equation

To gain some insight into the derivation of the error estimates, let us consider a discrete function $v_h \in H^1(J;V_h)$, and let us consider the following error decomposition for all $t \in J$:

$$e_h(t) := u_h(t) - v_h(t), \qquad \eta(t) := u(t) - v_h(t). \tag{66.9}$$

The conformity of the approximation setting implies that $b(u-u_h, y_h) = 0$ for all $y_h \in Y_h$, so that the discrete error $e_h \in X_h$ solves the parabolic problem $b(e_h, y_h) = b(\eta, y_h)$ for all $y_h \in Y_h$. This implies in particular that the following holds true in $L^2(J)$ for all $w_h \in V_h$:

$$(\partial_t e_h, w_h)_L + a(t; e_h, w_h) = \langle \partial_t \eta, w_h \rangle_{V',V} + a(t; \eta, w_h), \tag{66.10}$$

where we used the L-inner product for the time derivative of e_h. By using the same stability mechanisms as those invoked in the previous chapter in the continuous setting, the error equation (66.10) allows us to bound e_h in terms of η, and the error estimate then results from the triangle inequality. Thus, the only outstanding question is the choice of v_h to bound $\eta := u - v_h$.

To simplify some arguments, we henceforth assume that $u \in H^1(J;V)$ (this assumption requires that $u_0 \in V$). A simple, but somewhat naive, choice is to set $v_h(t) := \mathcal{I}_h(u(t))$ for all $t \in J$, where $\mathcal{I}_h : V \to V_h$ is any approximation operator having optimal approximation properties. This approach entails estimating the two terms composing the right-hand side of (66.10) by invoking the approximation properties of $\mathcal{I}_h$. Notice that we indeed have $\partial_t v_h \in L^2(J;V_h)$. This follows from $\partial_t v_h := \partial_t \mathcal{I}_h(u) = \mathcal{I}_h(\partial_t u)$, where the last equality results from Lemma 64.34 applied to the time-independent operator $\mathcal{I}_h$ (which is bounded on V) and the fact that $\partial_t u \in L^2(J;V)$ by assumption.

An alternative approach, which was introduced by Wheeler [285] for the heat equation, is to consider a suitable projection which relies on the differential operator in space. This idea will be reused in the context of the time-dependent Stokes equations (see §73.2.2) and of the time-dependent Friedrichs' systems (see §76.4.3). In the context of parabolic equations, we introduce the time-dependent projection $\Pi_h^{\mathrm{E}}(t) : V \to V_h$ s.t. for a.e. $t \in J$ and all $v \in V$, $\Pi_h^{\mathrm{E}}(t;v)$ is the unique solution to the following problem:

$$a(t; \Pi_h^{\mathrm{E}}(t;v), w_h) = a(t; v, w_h), \qquad \forall w_h \in V_h. \tag{66.11}$$

We henceforth abuse the language by saying that Π_h^{E} is an *elliptic projection* onto the finite element space V_h (see §32.4 where $a(v,w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)}$). Setting $e_h(t) := u_h(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ and $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$, and assuming that $\Pi_h^{\mathrm{E}}(\cdot;u(\cdot)) \in H^1(J;V_h)$, the error equation (66.10) becomes

$$(\partial_t e_h, w_h)_L + a(t; e_h, w_h) = \langle \partial_t \eta, w_h \rangle_{V',V}. \tag{66.12}$$

Thus, the crucial advantage of using an elliptic projection is that the right-hand side of (66.12) can be estimated without invoking $\|\eta\|_V$. One still needs to estimate $\partial_t \eta$ in weaker norms (e.g., the $\|\cdot\|_{V'}$-norm or the $\|\cdot\|_L$-norm). This can be done easily by invoking the approximation properties of the finite element setting if a is time-independent. If this is not the case, some mild additional assumptions on the time derivative of $a(t;\cdot,\cdot)$ are required. The first situation is addressed in §66.3.3 and the second situation in §66.3.4.

66.3.2 Basic error estimates

Let us start with an error estimate in the $L^2(J;V)$-norm.

Theorem 66.7 ($L^2(J;V)$-estimate). *Let $u \in X$ solve (66.3) and $u_h \in X_h$ solve (66.6). Assume $u \in H^1(J;V)$. Let $\eta(t) := u(t) - \mathcal{I}_h(u(t))$ for all $t \in J$, where $\mathcal{I}_h : V \to V_h$ is any approximation operator. The following holds true:*

$$\|u - u_h\|_{L^2(J;V)} \le \Big(1 + \frac{M}{\alpha}\Big)\|\eta\|_{L^2(J;V)} + \frac{1}{\alpha}\|\partial_t\eta\|_{L^2(J;V')} + \frac{1}{\sqrt{\alpha}}\|\eta(0)\|_L.$$

Proof. We consider the test function $w_h := e_h(t)$ for all $t \in J$ in the error equation (66.10). Invoking the coercivity of $a(t;\cdot,\cdot)$ and Young's inequality on the right-hand side, we infer that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_L^2 + \alpha\|e_h\|_V^2 \le \frac{1}{2\alpha}\|\partial_t\eta + A(\eta)\|_{V'}^2 + \frac{1}{2}\alpha\|e_h\|_V^2.$$

(Recall that $A(\eta)(t) = A(t)(\eta(t))$ for a.e. $t \in J$; see (65.6).) Rearranging the terms, integrating over $t \in J$, and dropping the nonnegative term $\|e_h(T)\|_L^2$ on the left-hand side gives

$$\alpha\|e_h\|_{L^2(J;V)}^2 \le \frac{1}{\alpha}\|\partial_t\eta + A(\eta)\|_{L^2(J;V')}^2 + \|e_h(0)\|_L^2.$$

(Notice that the above reasoning is the same as in the proof of Lemma 65.10.) Dividing by α, taking the square root, and since $\|\partial_t\eta + A(\eta)\|_{L^2(J;V')} \le \|\partial_t\eta\|_{L^2(J;V')} + M\|\eta\|_{L^2(J;V)}$, we infer that

$$\|e_h\|_{L^2(J;V)} \le \frac{1}{\alpha}\|\partial_t\eta\|_{L^2(J;V')} + \frac{M}{\alpha}\|\eta\|_{L^2(J;V)} + \frac{1}{\sqrt{\alpha}}\|e_h(0)\|_L. \tag{66.13}$$

The optimality property of $\mathcal{P}_{V_h}$ implies that $\|e_h(0)\|_L = \|\mathcal{P}_{V_h}(u_0 - v_h(0))\|_L \le \|u_0 - v_h(0)\|_L = \|\eta(0)\|_L$. Using this bound in (66.13) and invoking the triangle inequality for $u - u_h = \eta - e_h$ proves the assertion. □

Remark 66.8 (Supercloseness). Using the error equation (66.12), i.e., setting $e_h(t) := u_h(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ and $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$, and reasoning as in the above proof gives $\alpha\|e_h\|_{L^2(J;V)}^2 \le \frac{1}{\alpha}\|\partial_t\eta\|_{L^2(J;V')}^2 + \|e_h(0)\|_L^2$. This estimate exhibits a *supercloseness* phenomenon, i.e., the error on the left-hand side is measured in the V-norm, whereas the terms on the right-hand are measured in weaker norms. This property is central to the method often called in the literature *post-processing Galerkin* (see, e.g., García-Archilla et al. [134], García-Archilla and Titi [133]) and *nonlinear Galerkin methods* (see, e.g., Marion and Temam [224, 225], Guermond and Prudhomme [159, Rmk. 6.1]). See also Exercise 66.1. □

Let us now bound the error in the $C^0(\overline{J};L)$-norm. To this purpose, it is essential to avoid invoking $\|\eta\|_V$ and this is the reason why we consider the elliptic projection defined in (66.11). To simplify some arguments, we assume that the bilinear form a is time-independent (we shall return to the general setting in §66.3.4). Then the *elliptic projection* $\Pi_h^{\mathrm{E}} : V \to V_h$ is time-independent and is such that for all $v \in V$,

$$a(\Pi_h^{\mathrm{E}}(v), w_h) = a(v, w_h), \qquad \forall w_h \in V_h. \tag{66.14}$$

We introduce the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, i.e., the smallest constant s.t. $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$.

Theorem 66.9 ($C^0(\overline{J};L)$-estimate). *Let $u \in X$ solve (66.3) and $u_h \in X_h$ solve (66.6). Assume $u \in H^1(J;V)$. Assume that the bilinear form a is time-independent. Letting $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$ for all $t \in (0,T]$, we have, with $J_t := (0,t)$,*

$$\|(u - u_h)(t)\|_L \le \|\eta(t)\|_L + \frac{1}{\sqrt{\alpha}}\|e^{-\frac{t-\cdot}{\rho}}\partial_t\eta\|_{L^2(J_t;V')} + e^{-\frac{t}{\rho}}\|\eta(0)\|_L. \tag{66.15}$$

Proof. See Exercise 66.3. □

Remark 66.10 (Exponential decay). Similarly to the a priori bound established in Lemma 65.11 in the continuous setting, the error estimate (66.15) shows that the error in the L-norm induced by the approximation of the initial condition u_0 decays exponentially fast with time. □

Remark 66.11 (Bounding the $\|\cdot\|_{V'}$-norm). Since the duality pairing between V' and V is an extension of the L-inner product, we infer that $\|\phi\|_{V'} \le \iota_{L,V}\|\partial_t\phi\|_L$ for all $\phi \in L$. Applying this bound to $\partial_t\eta$, the error estimates from Theorem 66.7 and Theorem 66.9 become

$$\|u-u_h\|_{L^2(J;V)} \le \Big(1+\frac{M}{\alpha}\Big)\|\eta\|_{L^2(J;V)} + \sqrt{\frac{\rho}{2\alpha}}\|\partial_t\eta\|_{L^2(J;L)} + \frac{1}{\sqrt{\alpha}}\|\eta(0)\|_L,$$

and

$$\|(u-u_h)(t)\|_L \le \|\eta(t)\|_L + \sqrt{\frac{\rho}{2}}\|e^{-\frac{t-\cdot}{\rho}}\partial_t\eta\|_{L^2(J_t;L)} + e^{-\frac{t}{\rho}}\|\eta(0)\|_L, \tag{66.16}$$

where we used that $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$. □

66.3.3 Application to the heat equation

We now illustrate the above error estimates on the heat equation, where $V := H^1_0(D)$, $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$, $L := L^2(D)$, and $V' = H^{-1}(D)$ (see Example 66.1). The time scale becomes $\rho := \frac{2}{C_{\rm PS}^2}\frac{\ell_D^2}{\alpha}$ since $\iota_{L,V} = \frac{C_{\rm PS}}{\ell_D}$ owing to the Poincaré–Steklov inequality. The discretization in space relies on continuous finite elements, i.e., we take $V_h := \mathcal{P}^{\rm g}_{k,0}(\mathcal{T}_h) \subset H^1_0(D)$ (see §19.4).

Corollary 66.12 ($L^2(J;V)$-estimate, heat equation). *Let $r \in [1,k]$, where $k \ge 1$ is the degree of the finite elements used to build the discrete space V_h. Assume that $u \in L^2(J;H^{r+1}(D)) \cap H^1(J;H^r(D))$ (so that $u_0 \in H^r(D)$). There is c s.t. for all $h \in \mathcal{H}$, α, and M,*

$$\|u-u_h\|^2_{L^2(J;H^1_0(D))} \le c\Bigg(\sum_{K\in\mathcal{T}_h} h_K^{2r}\Big(\frac{M^2}{\alpha^2}\int_J |u(t)|^2_{H^{r+1}(K)}\mathrm{d}t + C_{\rm PS}^2\frac{\rho^2}{\ell_D^2}\int_J|\partial_t u(t)|^2_{H^r(K)}\mathrm{d}t + \frac{1}{\alpha}|u_0|^2_{H^r(K)}\Big)\Bigg). \tag{66.17}$$

Proof. We invoke Theorem 66.7 (actually the bound from Remark 66.11) with $v_h(t) := \mathcal{I}^{\rm g,av}_{h0}(u(t))$ for a.e. $t \in J$, where $\mathcal{I}^{\rm g,av}_{h0}$ is the quasi-interpolation operator with zero boundary trace constructed in §22.4.2. Owing to the definition of the time scale ρ, we infer that

$$\|u-u_h\|_{L^2(J;H^1_0(D))} \le \Big(1+\frac{M}{\alpha}\Big)\|u-\mathcal{I}^{\rm g,av}_{h0}(u)\|_{L^2(J;H^1_0(D))} + \frac{C_{\rm PS}}{2}\frac{\rho}{\ell_D}\|\partial_t(u-\mathcal{I}^{\rm g,av}_{h0}(u))\|_{L^2(J;L^2(D))} + \frac{1}{\sqrt{\alpha}}\|u_0-\mathcal{I}^{\rm g,av}_{h0}(u_0)\|_{L^2(D)}.$$

The estimate (66.17) follows from Theorem 22.14 once we observe that $\partial_t(\mathcal{I}^{\rm g,av}_{h0}(u)) = \mathcal{I}^{\rm g,av}_{h0}(\partial_t u)$ owing to Lemma 64.34, $\partial_t u \in L^2(J;L^2(D))$ by assumption, and that $\mathcal{I}^{\rm g,av}_{h0}$ is bounded in $L^2(D)$. □

Corollary 66.13 (Improved $C^0(\overline{J};L^2(D))$-estimate, heat equation). *Let $r \in [1,k]$, where $k \ge 1$ is the degree of the finite elements used to build the discrete space V_h. Assume that the diffusion coefficient κ is time-independent. Assume that there is some elliptic regularity pickup*

in the adjoint problem, i.e., there are $s \in (0,1]$ and $c_{\mathrm{smo}} > 0$ s.t. for all $g \in L^2(D)$, the unique solution $\xi_g \in V$ s.t. $a(v,\xi_g) = (g,v)_{L^2(D)}$ for all $v \in V$ satisfies $\|\xi_g\|_{H^{1+s}(D)} \le c_{\mathrm{smo}}\alpha^{-1}\ell_D^2\|g\|_{L^2(D)}$. Assume that $u \in H^1(J;H^{r+1}(D))$. There is c, proportional to $(\frac{M}{\alpha})^2$, s.t. the following holds true for all $h \in \mathcal{H}$ and all $t \in (0,T]$:

$$\|(u-u_h)(t)\|_{L^2(D)} \le c\,h^s\ell_D^{1-s}\Big(\sum_{K\in\mathcal{T}_h} h_K^{2r}\Big(|u(t)|^2_{H^{r+1}(K)} + e^{-2\frac{t}{\rho}}|u_0|^2_{H^{r+1}(K)} + \rho\int_0^t e^{-2\frac{t-\tau}{\rho}}|\partial_t u(\tau)|^2_{H^{r+1}(K)}\,\mathrm{d}\tau\Big)\Big)^{\frac12}. \tag{66.18}$$

Proof. We invoke Theorem 66.9 (actually the bound (66.16)). Notice that the smoothness assumption on u implies that $u \in C^0(\overline{J};H^{r+1}(D))$. Owing to the elliptic regularity pickup of the adjoint problem, and adapting the proof of Lemma 32.11 and Theorem 32.15, we infer that for all $v \in H^{r+1}(D)\cap H_0^1(D)$,

$$\|v-\Pi_h^{\mathrm{E}}(v)\|_{L^2(D)} \le c_1\,h^s\ell_D^{1-s}|v-\Pi_h^{\mathrm{E}}(v)|_{H^1(D)} \le c_2\,h^s\ell_D^{1-s}\inf_{w_h\in V_h}|v-w_h|_{H^1(D)} \le c_3\,h^s\ell_D^{1-s}\Big(\sum_{K\in\mathcal{T}_h}h_K^{2r}|v|^2_{H^{r+1}(K)}\Big)^{\frac12},$$

where the last bound follows from the approximation properties of finite elements. Here, c_1 is proportional to $\frac{M}{\alpha}$ and c_2, c_3 are proportional to $(\frac{M}{\alpha})^2$. We conclude by combining the above two bounds and observing that $\partial_t\eta = \partial_t(u-\Pi_h^{\mathrm{E}}(u)) = \partial_t u - \Pi_h^{\mathrm{E}}(\partial_t u)$ since a is time-independent. □

Remark 66.14 (Decay rate). If $u \in L^2(J;H^{r+1}(D))\cap H^1(J;H^r(D))$, the estimate (66.17) shows that $\|u-u_h\|_{L^2(J;H_0^1(D))}$ converges with the rate $\mathcal{O}(h^r)$, $r \in [1,k]$. Moreover, assuming $u \in H^1(J;H^{r+1}(D))$, the estimate (66.18) shows that the error $\|(u-u_h)(t)\|_L$ converges for all times $t \in (0,T]$ with the quasi-optimal rate $\mathcal{O}(h^{r+s})$, $r \in [1,k]$. The convergence rate takes the optimal value $\mathcal{O}(h^{r+1})$ if there is full elliptic regularity pickup for the adjoint problem ($s=1$). Notice in passing that using the error equation (66.10) leads to a bound on $\|u-u_h\|_{C^0(\overline{J};L^2(D))}$ with the suboptimal decay rate $\mathcal{O}(h^r)$; see Exercise 66.2. □

Remark 66.15 (Smoothness of $\partial_t u$). In Corollary 66.12, the smoothness assumption on the time derivative can be relaxed to $u \in H^1(J;H^{r-1}(D))$, but to do so one needs to consider an interpolant with superconvergent approximation properties in $H^{-1}(D)$. One possibility is to use a variant of the Scott–Zhang interpolation operator preserving mean-values over element patches and boundary conditions, as done in Tantardini and Veeser [270, p. 337]. This operator, say $\mathcal{I}_{h0}^{\mathrm{TV}}$, is such that $\|v-\mathcal{I}_{h0}^{\mathrm{TV}}(v)\|^2_{H^m(D)} \le c\sum_{K\in\mathcal{T}_h}h_K^{2(s-m)}|v|^2_{H^s(K)}$ for all $m \in \{-1,0,1\}$ and $\max(0,m) \le s \le k+1$. Using this operator leads to the bound

$$\|u-u_h\|^2_{L^2(J;H_0^1(D))} \le c\Big(\sum_{K\in\mathcal{T}_h}h_K^{2r}\Big(\frac{M^2}{\alpha^2}\int_J|u(t)|^2_{H^{r+1}(K)}\mathrm{d}t + C_{\mathrm{PS}}^4\frac{\rho^2}{\ell_D^4}\int_J|\partial_t u(t)|^2_{H^{r-1}(K)}\mathrm{d}t + \frac{1}{\alpha}|u_0|^2_{H^r(K)}\Big)\Big).$$

If one is willing to accept the loss of localization in the error estimate, one can also use the L^2-orthogonal projection, since $\|v-\mathcal{P}_{h0}(v)\|_{H^{-1}(D)} \le ch\|v-\mathcal{P}_{h0}(v)\|_{L^2(D)}$; see Exercise 22.6 and Remark 22.23. On the other hand it is not possible in Corollary 66.13 to lower the smoothness requirement on the time derivative to $\partial_t u \in L^2(J;H^r(D))$, since the elliptic projection does not have superconvergent approximation properties in $H^{-1}(D)$. □

66.3.4 Extension to time-varying diffusion

The above analysis based on the elliptic projection can be extended under reasonable assumptions to the case where the bilinear form a is time-dependent. In the context of the heat equation, this means that the diffusion coefficient κ is time-dependent. Recall that the time-dependent elliptic projection $\Pi_h^{\mathrm{E}}(t) \in \mathcal{L}(V;V_h)$ is defined in (66.11) for a.e. $t \in J$. We assume that $A : J \to \mathcal{L}(V;V')$ is strongly differentiable and there is M' such that $\|\partial_t A(t)(v)\|_{V'} \le \rho^{-1} M' \|v\|_V$ for all $v \in V$ and a.e. $t \in J$. Here, we used the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$ so that the constants M and M' have the same units. We define the bilinear form $\dot a(t;v,w) := \langle \partial_t A(t)(v), w\rangle_{V',V}$ for all $v,w \in V$ and a.e. $t \in J$. For the heat equation, $M' := \rho\|\partial_t\kappa\|_{L^\infty(D\times J)}$ and $\dot a(t;v,w) := \int_D \partial_t\kappa(x,t)\nabla v(x)\cdot\nabla w(x)\,\mathrm{d}x$.

Lemma 66.16 (Commuting with time derivative). *Assume that $u \in H^1(J;V)$. Then the function $J \ni t \mapsto \Pi_h^{\mathrm{E}}(t;u(t)) \in V_h$ is in $H^1(J;V_h)$, and we have for all $w_h \in V_h$ and a.e. $t \in J$,*

$$a(t;\partial_t(\Pi_h^{\mathrm{E}}(t;u(t))) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t)), w_h) = \dot a(t;u(t) - \Pi_h^{\mathrm{E}}(t;u(t)), w_h). \tag{66.19}$$

Proof. We first establish (66.19) for smooth functions. We apply (66.11) with $v(t) \in C^\infty(\overline{J};V)$, differentiate this relation in time, and use the definition of $\Pi_h^{\mathrm{E}}(t;\partial_t v(t))$. The coercivity of a together with the boundedness of a and $\dot a$ shows that there is c s.t. $\|\Pi_h^{\mathrm{E}}(t;v(t))\|_{H^1(J;V_h)} \le c\|v\|_{H^1(J;V)}$ for all $v \in H^1(J;V)$ and all $h \in \mathcal{H}$ (see Exercise 66.4(iii)). We conclude by invoking the density of $C^\infty(\overline{J};V)$ in $H^1(J;V)$ (see Theorem 64.36 with $V = W$, $p = q := 2$) and the linearity of the map $H^1(J;V) \ni v \mapsto \Pi_h^{\mathrm{E}}(\cdot;v(\cdot)) \in H^1(J;V_h)$. □

Lemma 66.16 implies that the function $t \mapsto \eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ is in $H^1(J;V)$ (recall that $u \in H^1(J;V)$ by assumption). We can then apply the estimate (66.16) with this function. To derive an error estimate, it remains to establish approximation properties of η and $\partial_t\eta$ in L. For simplicity, we focus on the functional setting of the heat equation where $L := L^2(D)$ so that we can invoke the elliptic regularity theory. The time dependence of Π_h^{E} is irrelevant to estimate $\|\eta\|_{L^2(D)}$, but it makes estimating $\|\partial_t\eta\|_{L^2(D)}$ more delicate.

Lemma 66.17 (Estimate on $\|\partial_t\eta(t)\|_{L^2(D)}$). *Assume that the elliptic regularity pickup from Corollary 66.13 holds true with $s \in (\frac12, 1]$. Assume that κ is continuously differentiable in time and set $M' := \rho\|\partial_t\kappa\|_{L^\infty(D\times J)}$. Assume further that $\partial_t\kappa$ is smooth enough so that*

$$|\dot a(t;w,z)| \le \rho^{-1} M'' |w|_{H^{1-s}(D)} |z|_{H^{1+s}(D)}, \quad \forall w,z \in H_0^1(D). \tag{66.20}$$

Letting $c_\kappa := \big(1+\frac{M}{\alpha}\big)\frac{M'}{\alpha} + \big(\frac{M}{\alpha}\big)^s\frac{M''}{\alpha}$, there is c s.t. for all $h \in \mathcal{H}$, α, M, M', M'', and a.e. $t \in J$,

$$\begin{aligned}\|\partial_t\eta(t)\|_{L^2(D)} \le{}& \|\partial_t u(t) - \Pi_h^{\mathrm{E}}(t;\partial_t u(t))\|_{L^2(D)} \\ &+ c\,\rho^{-1} c_\kappa h^{s^2} \ell_D^{1-s^2} |u(t) - \Pi_h^{\mathrm{E}}(t;u(t))|_{H^1(D)}.\end{aligned} \tag{66.21}$$

Proof. See Exercise 66.4. □

Remark 66.18 (Lemma 66.17). The estimate (66.21) is somewhat suboptimal since $\big(\frac{h}{\ell_D}\big)^{s^2} \le \big(\frac{h}{\ell_D}\big)^s \le 1$ (because $h \le \ell_D$ and $s \le 1$). Optimality is recovered if full elliptic regularity pickup holds true, i.e., if $s = 1$. Furthermore, assuming that κ is time-independent on the boundary (so that $(\partial_t\kappa)_{|\partial D} = 0$) and that $\partial_t\kappa$ satisfies the multiplier property $\|\partial_t\kappa\nabla z\|_{\boldsymbol{H}_0^s(D)} \le \rho^{-1}M''\|\nabla z\|_{\boldsymbol{H}^s(D)}$, the hypothesis (66.20) follows from $|\dot a(t;w,z)| \le \|\nabla w\|_{\boldsymbol{H}^{-s}(D)}\|\partial_t\kappa\nabla z\|_{\boldsymbol{H}_0^s(D)}$. □

Corollary 66.19 (Heat equation with time-varying diffusion). *Let $r \in [1,k]$, where $k \geq 1$ is the degree of the finite elements used to construct the discrete space V_h. Under the assumptions of Lemma 66.17, if $u \in H^1(J;H^{r+1}(D))$, there is c, proportional to $c_\kappa \frac{M}{\alpha}$, s.t. for all $h \in \mathcal{H}$ and all $t \in (0,T]$,*

$$\|(u-u_h)(t)\|_{L^2(D)} \leq c\, h^{s^2} \ell_D^{1-s^2} \Bigg(\sum_{K \in \mathcal{T}_h} h_K^{2r} \Big(|u(t)|^2_{H^{r+1}(K)} + e^{-2\frac{t}{\rho}} |u_0|^2_{H^{r+1}(K)} + \int_0^t e^{-2\frac{t-\xi}{\rho}} \big(\rho |\partial_\xi u(\xi)|^2_{H^{r+1}(K)} + \rho^{-1} |u(\xi)|^2_{H^{r+1}(K)} \big) \,\mathrm{d}\xi \Big) \Bigg)^{\frac{1}{2}}. \tag{66.22}$$

Proof. The only difference with respect to the proof of Corollary 66.13 lies in the bound of the term $\rho \int_0^t e^{-2\frac{t-\xi}{\rho}} \|\partial_\xi \eta(\xi)\|^2_{L^2(D)} \,\mathrm{d}\xi$ from (66.16), which we now estimate by means of Lemma 66.17. Let $\xi \in J_t := (0,t)$. The approximation properties of $\Pi_h^{\mathrm{E}}(\xi)$ imply that

$$\|\partial_\xi u(\xi) - \Pi_h^{\mathrm{E}}(\xi; \partial_\xi u(\xi))\|_{L^2(D)} \leq c_1 h^s \ell_D^{1-s} \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |\partial_\xi u(\xi)|^2_{H^{r+1}(K)} \right)^{\frac{1}{2}}$$

with c_1 proportional to $\left(\frac{M}{\alpha}\right)^2$, and

$$|u(\xi) - \Pi_h^{\mathrm{E}}(\xi; u(\xi))|_{H^1(D)} \leq c_2 \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u(\xi)|^2_{H^{r+1}(K)} \right)^{\frac{1}{2}}$$

with c_2 proportional to $\frac{M}{\alpha}$. Since $c_\kappa \frac{M}{\alpha} \geq \left(\frac{M}{\alpha}\right)^2$ and $\left(\frac{h}{\ell_D}\right)^{s^2} \leq \left(\frac{h}{\ell_D}\right)^s \leq 1$ since $s \in (0,1]$, inserting the above bounds into (66.16) proves the assertion. □

Exercises

Exercise 66.1 ($L^2(J;V)$-estimate using elliptic projection). Use the notation from §66.3.1. Assume that the elliptic projection is time-independent and set $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$ for all $t \in J$. Prove that

$$\|u - u_h\|_{L^2(J;V)} \leq \|\eta\|_{L^2(J;V)} + \frac{1}{\alpha} \|\partial_t \eta\|_{L^2(J;V')} + \frac{2}{\sqrt{\alpha}} \|\eta(0)\|_L.$$

(*Hint*: use the error equation (66.12).)

Exercise 66.2 (Naive $C^0(\overline{J};L)$-estimate). Use the proof of Theorem 66.7 to derive an upper bound on $\|u - u_h\|_{C^0(\overline{J};L)}$. (*Hint*: integrate (66.10) in time over the interval $J_s := (0,s)$ for all $s \in (0,T]$.) Assuming smoothness, is the convergence rate of this error estimate optimal for the heat equation? What is the term that limits the convergence rate?

Exercise 66.3 (Theorem 66.9). Prove the error estimate (66.15). (*Hint*: see Exercise 65.4.)

Exercise 66.4 (Lemma 66.17). Let $\Pi_h^{\mathrm{E}}(t) \in \mathcal{L}(H_0^1(D);V_h)$ be defined in (66.11) for the time-dependent heat equation. Let $u \in H^1(J;H_0^1(D))$ and set $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$ for a.e. $t \in J$. (i) Prove that

$$|\partial_t \eta(t)|_{H^1(D)} \leq |\partial_t u(t) - \Pi_h^{\mathrm{E}}(t; \partial_t u(t))|_{H^1(D)} + \rho^{-1} \frac{M'}{\alpha} |\eta(t)|_{H^1(D)}.$$

(ii) Prove (66.21). (*Hint*: use the adjoint problem $a(t; v, \xi(t)) = (\delta_h(t), v)_{L^2(D)}$ for all $v \in H_0^1(D)$, with $\delta_h(t) := \partial_t(\Pi_h^{\mathrm{E}}(t; u(t))) - \Pi_h^{\mathrm{E}}(t; \partial_t u(t))$ for a.e. $t \in J$, and show that

$$\|\delta_h(t)\|_{L^2(D)}^2 = a(t; \delta_h(t), \xi(t) - w_h) + \dot{a}(t; \eta(t), w_h - \xi(t)) + \dot{a}(t; \eta(t), \xi(t)),$$

for all $w_h \in V_h$.) (iii) Show that $\|\Pi_h^{\mathrm{E}}(t; u(t))\|_{H^1(J;V_h)} \leq c(\alpha, M, \frac{M'}{\rho})\|u\|_{H^1(J;V)}$ for all $u \in C^\infty(\overline{J}; V)$ and all $h \in \mathcal{H}$.

Chapter 67

Implicit and explicit Euler schemes

In the previous chapter, we studied the space semi-discrete parabolic problem (66.6). The goal is now to discretize (66.6) in time. Since this problem is a system of coupled (linear) ODEs, its time discretization can be done by using one of the numerous time-stepping techniques available from the literature. In this chapter, we focus on the implicit (or backward) Euler scheme and on the explicit (or forward) Euler scheme, which are both first-order accurate in time. Second-order implicit schemes called BDF2 and Crank–Nicolson are investigated in Chapter 68. The standard viewpoint in the literature is to interpret the above schemes as finite differences in time. This is the perspective we adopt in this chapter and the next one. We broaden the perspective in Chapters 69 and 70 by introducing a discrete space-time formulation and by considering higher-order time discretization methods.

67.1 Implicit Euler scheme

One of the most basic methods to discretize in time the semi-discrete problem (66.6) is the *implicit Euler scheme*. We analyze this method in this section by adopting the finite difference viewpoint.

67.1.1 Time mesh

Let $N > 0$ be a positive natural number. We divide the time interval $J := (0, T)$ with $T > 0$ into N subintervals J_n for all $n \in \mathcal{N}_\tau := \{1{:}N\}$. All the intervals are of equal length to simplify the notation (this is not a theoretical requirement), i.e., we define the *time step* to be $\tau := \frac{T}{N}$, the *discrete time nodes* to be $t_n := n\tau$, for all $n \in \overline{\mathcal{N}}_\tau := \{0{:}N\}$, and we set $J_n := (t_{n-1}, t_n]$ for all $n \in \mathcal{N}_\tau$, so that $\overline{J} = \bigcup_{n\in\mathcal{N}_\tau} \overline{J}_n$.

Given a Banach space B with norm $\|\cdot\|_B$ and seminorm $|\cdot|_B$, and a collection of members of B, say $v_\tau := (v^n)_{n\in\mathcal{N}_\tau} \in B^N$, where $v^n \in B$ is associated with the time node t_n, we define the

time-discrete norms and seminorms

$$\|v_\tau\|^2_{\ell^2(J;B)} := \sum_{n\in\mathcal{N}_\tau} \tau\|v^n\|_B^2, \qquad |v_\tau|^2_{\ell^2(J;B)} := \sum_{n\in\mathcal{N}_\tau} \tau|v^n|_B^2, \tag{67.1a}$$

$$\|v_\tau\|_{\ell^\infty(\overline{J};B)} := \max_{n\in\overline{\mathcal{N}}_\tau} \|v^n\|_B, \qquad |v_\tau|_{\ell^\infty(\overline{J};B)} := \max_{n\in\overline{\mathcal{N}}_\tau} |v^n|_B. \tag{67.1b}$$

One should think of $\|v_\tau\|_{\ell^2(J;B)}$ and $\|v_\tau\|_{\ell^\infty(\overline{J};B)}$ as the time-discrete counterparts of $\|v\|_{L^2(J;B)}$ and $\|v\|_{C^0(\overline{J};B)}$, respectively. These norms and seminorms will be useful to state the stability results and the error estimates.

67.1.2 Principle and algebraic realization

Recall that the model parabolic problem is posed using the trial space $X := \{v \in L^2(J;V) \mid \partial_t v \in L^2(J;V')\}$ and the test space $Y := L\times L^2(J;V)$, where $(V, L \equiv L', V')$ is a Gelfand triple. Let $(V_h)_{h\in\mathcal{H}}$ be a sequence of finite-dimensional subspaces of V which are constructed using a mesh $\mathcal{T}_h$ and a reference finite element (see §19.2.1 for H^1-conforming subspaces). In the entire chapter, we assume that the same mesh $\mathcal{T}_h$ is used at all times; see Remark 67.2. The semi-discretization in space uses the semi-discrete trial space $X_h := H^1(J;V_h) \subset X$ and the semi-discrete test space $Y_h := V_h\times L^2(J;V_h) \subset Y$. Our starting point is the semi-discrete formulation (66.6): Find $u_h \in X_h$ s.t.

$$(\partial_t u_h(t), w_h)_L + a(t;u_h(t),w_h) = \langle f(t), w_h\rangle_{V',V}, \tag{67.2a}$$

$$u_h(0) = \mathcal{P}_{V_h}(u_0), \tag{67.2b}$$

where (67.2a) holds in $L^2(J)$ for all $w_h \in V_h$, and where $\mathcal{P}_{V_h} : L \to V_h$ is the L-orthogonal projection onto V_h. To avoid technicalities with point values in time, we are going to assume that $f \in C^0(\overline{J};V')$ and that the map $J \ni t \to a(t;v,w) \in \mathbb{R}$ is continuous for all $t \in \overline{J}$ and all $v, w \in V$. As a result, we have $u_h \in C^1(\overline{J};V_h)$.

The main idea is to consider the ODEs in (67.2) at the discrete time nodes $(t_n)_{n\in\mathcal{N}_\tau}$ and use the backward first-order finite difference formula to approximate the time derivative as $\partial_t u_h(t_n) = \frac{u_h(t_n)-u_h(t_{n-1})}{\tau} + \mathcal{O}(\tau)$. Multiplying this approximation by τ and setting $u_h^0 := u_h(0) = \mathcal{P}_{V_h}(u_0)$, the discrete problem consists of seeking a sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ s.t. for all $n \in \mathcal{N}_\tau$,

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a^n(u_h^n, w_h) = \tau\langle f^n, w_h\rangle_{V',V}, \quad \forall w_h \in V_h, \tag{67.3}$$

where $a^n(\cdot,\cdot) := a(t_n;\cdot,\cdot)$ and $f^n := f(t_n) \in V'$.

Let $I := \dim(V_h)$ and $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h (e.g., the global shape functions in V_h). Let $\mathsf{U}^n \in \mathbb{R}^I$ be the coordinate vector of u_h^n in this basis for all $n \in \overline{\mathcal{N}}_\tau$, i.e., $u_h^n(\boldsymbol{x}) := \sum_{i\in\{1:I\}} \mathsf{U}_i^n\varphi_i(\boldsymbol{x})$. Recall that the *stiffness matrix* $\mathcal{A}(t) \in \mathbb{R}^{I\times I}$ and the *mass matrix* $\mathcal{M} \in \mathbb{R}^{I\times I}$ are defined s.t.

$$\mathcal{A}_{ij}(t) := a(t;\varphi_j,\varphi_i), \qquad \mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L, \qquad \forall i,j \in \{1{:}I\}. \tag{67.4}$$

The mass matrix is symmetric positive definite, and the stiffness matrix is positive definite (see §28.2.2-§28.2.3). Using the above notation, (67.3) is recast as follows: For all $n \in \mathcal{N}_\tau$, find $\mathsf{U}^n \in \mathbb{R}^I$ s.t.

$$\mathcal{M}(\mathsf{U}^n - \mathsf{U}^{n-1}) + \tau\mathcal{A}^n\mathsf{U}^n = \tau\mathsf{F}^n, \tag{67.5}$$

with $\mathcal{A}^n := \mathcal{A}(t_n)$ and the components of $\mathsf{F}^n \in \mathbb{R}^I$ are $(\langle f^n, \varphi_i\rangle_{V',V})_{i\in\{1:I\}}$. Rearranging the terms in (67.5) gives

$$(\mathcal{M} + \tau\mathcal{A}^n)\mathsf{U}^n = \mathcal{M}\mathsf{U}^{n-1} + \tau\mathsf{F}^n, \tag{67.6}$$

showing that each step of the implicit Euler scheme entails solving a linear system with the positive definite matrix $\mathcal{M} + \tau\mathcal{A}^n$.

Remark 67.1 (Variants). If the source term f or the bilinear form a do not have point values in time, it is possible to consider averaged values over the time subintervals, e.g., one can set $a^n(\cdot,\cdot) := \frac{1}{\tau}\int_{J_n} a(t;\cdot,\cdot)\mathrm{d}t$ and $f^n := \frac{1}{\tau}\int_{J_n} f(t)\mathrm{d}t$ in (67.3). Several choices of the initial condition are also possible as long as u_h^0 optimally approximates u_0. □

Remark 67.2 (Time-dependent meshes). Considering time-dependent meshes is possible, but the analysis of the time-stepping schemes becomes more intricate. Moreover, one must bear in mind that changing the mesh too frequently can be problematic, even in simple problems as the one-dimensional heat equation. A counterexample by Dupont [113] shows that frequent mesh changes can introduce excessive dissipation and hamper convergence. □

67.1.3 Stability

The stability mechanism that comes into play in the analysis of the implicit Euler scheme is the same as for the continuous and the semi-discrete problems. Recall that α and M denote the coercivity and the boundedness constants associated with the bilinear form a. To allow for a more compact notation, we define the sequence of approximate time derivatives $\delta_\tau u_{h\tau} \in (V_h)^N$ s.t. $(\delta_\tau u_{h\tau})^n := \frac{1}{\tau}(u_h^n - u_h^{n-1})$ for all $n \in \mathcal{N}_\tau$.

Lemma 67.3 ($\ell^2(J;V)$-stability). *Let $u_{h\tau} \in (V_h)^N$ solve (67.3) with the sequence of source terms $f_\tau := (f^n)_{n\in\mathcal{N}_\tau} \in (V')^N$. The following holds true:*

$$\alpha\|u_{h\tau}\|_{\ell^2(J;V)}^2 + \tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2 + \|u_h^N\|_L^2 \le \frac{1}{\alpha}\|f_\tau\|_{\ell^2(J;V')}^2 + \|u_h^0\|_L^2. \tag{67.7}$$

Proof. Using $w_h := u_h^n$ as the test function in (67.3) leads to

$$(u_h^n - u_h^{n-1}, u_h^n)_L + \tau a^n(u_h^n, u_h^n) = \tau\langle f^n, u_h^n\rangle_{V',V}. \tag{67.8}$$

The key stability mechanism for the time derivative hinges on the identity

$$(u_h^n - u_h^{n-1}, u_h^n)_L = \frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 + \frac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2. \tag{67.9}$$

Owing to this identity, the coercivity of a^n, and bounding the right-hand side of (67.8) by using Young's inequality, we obtain

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \|u_h^n - u_h^{n-1}\|_L^2 + \alpha\tau\|u_h^n\|_V^2 \le \frac{1}{\alpha}\tau\|f^n\|_{V'}^2. \tag{67.10}$$

Summing over $n \in \mathcal{N}_\tau$, exploiting the telescoping form of the first two terms on the left-hand side, and rearranging the terms proves the assertion. □

Remark 67.4 (Comparison). The stability estimate (67.7) has the same structure as the a priori estimate derived in Lemma 65.10 for the continuous problem. Actually, both proofs use the same arguments, the only difference being that the integration by parts in time is replaced by the identity (67.9) and a summation over $n \in \mathcal{N}_\tau$. The additional bound on $\tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2$ in (67.7) provides some (weak) control on the time derivative. □

As in the previous chapters, we now derive a sharper stability estimate that captures the exponentially decaying influence of the initial data. We consider the $\ell^\infty(\overline{J};L)$-norm instead of the $C^0(\overline{J};L)$-norm since the setting is discrete in time. Recall that $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, i.e., it is the smallest constant s.t. $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$. Define the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$.

Lemma 67.5 ($\ell^\infty(\overline{J};L)$-stability, exponential decay). *Let $u_{h\tau} \in (V_h)^N$ solve (67.3) with $f_\tau := (f^n)_{n\in\mathcal{N}_\tau} \in (V')^N$. Assume (for simplicity) that $\tau \le \frac{1}{2}\rho$. The following holds true for all $n \in \mathcal{N}_\tau$:*

$$\|u_h^n\|_L^2 \le e^{-\frac{t_n}{\rho}}\|u_h^0\|_L^2 + \frac{1}{\alpha}\sum_{k\in\{1:n\}} \tau e^{-\frac{t_n - t_{k-1}}{\rho}}\|f^k\|_{V'}^2. \tag{67.11}$$

Proof. Using the stability estimate (67.10), we infer that

$$\begin{aligned}\Big(1+2\frac{\tau}{\rho}\Big)\|u_h^n\|_L^2 &\le \|u_h^n\|_L^2 + \alpha\tau\iota_{L,V}^{-2}\|u_h^n\|_L^2\\ &\le \|u_h^n\|_L^2 + \alpha\tau\|u_h^n\|_V^2 \le \|u_h^{n-1}\|_L^2 + \frac{1}{\alpha}\tau\|f^n\|_{V'}^2.\end{aligned}$$

Applying the incremental Gronwall lemma from Exercise 67.1 with $\gamma := 2\frac{\tau}{\rho}$, $a_n := \|u_h^n\|_L^2$, and $b_n := \frac{1}{\alpha}\tau\|f^n\|_{V'}^2$ yields

$$\|u_h^n\|_L^2 \le \frac{\|u_h^0\|_L^2}{(1+2\frac{\tau}{\rho})^n} + \frac{1}{\alpha}\tau\sum_{k\in\{1:n\}}\frac{\|f^k\|_{V'}^2}{(1+2\frac{\tau}{\rho})^{n-k+1}}.$$

Since $2\frac{\tau}{\rho} \le 1$ by assumption, we have $(1+2\frac{\tau}{\rho})^{-1} \le e^{-\frac{\tau}{\rho}}$. The bound (67.11) follows readily. □

67.1.4 Error analysis

Let us start by estimating the error in the $\ell^2(J;V)$-norm. Recalling the discussion in §66.3 we first write the error equation by using a generic operator $\mathcal{I}_h : V \to V_h$ having optimal approximation properties, and then we make particular choices. We consider the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$.

Theorem 67.6 ($\ell^2(J;V)$-estimate). *Let u solve the model parabolic problem (65.10) and assume that $u \in H^1(J;V)\cap H^2(J;V')$. Let $\eta(t) := u(t) - \mathcal{I}_h(u(t))$ for all $t \in J$. Set $u_\tau := (u(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$ and $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$. Let $u_{h\tau} \in (V_h)^N$ solve (67.3). Then we have*

$$\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le c_1\tau\|\partial_{tt}u\|_{L^2(J;V')} + c_2\|\eta_\tau\|_{\ell^2(J;V)} + c_1\|\partial_t\eta\|_{L^2(J;V')} + c_3\|\eta(0)\|_L, \tag{67.12}$$

with $c_1 := \frac{\sqrt{3}}{\alpha}$, $c_2 := 1 + \frac{\sqrt{3}M}{\alpha}$, and $c_3 := \frac{1}{\sqrt{\alpha}}$.

Proof. (1) For all $n \in \mathcal{N}_\tau$, we set $e_h^n := u_h^n - \mathcal{I}_h(u(t_n)) \in V_h$, leading to the error decomposition $u_h^n - u(t_n) = e_h^n - \eta(t_n)$. The idea is to estimate $(e_h^n)_{n\in\mathcal{N}_\tau}$ in terms of η by invoking the stability estimate (67.7), and conclude by means of the triangle inequality.
(2) We observe that for all $n \in \mathcal{N}_\tau$ and all $w_h \in V_h$,

$$(u(t_n) - u(t_{n-1}), w_h)_L + \tau a^n(u(t_n), w_h) = \tau\langle f^n + \psi^n, w_h\rangle_{V',V},$$

(recall that $H^1(J;V) \hookrightarrow C^0(\overline{J};V)$) with

$$\psi^n := \frac{u(t_n) - u(t_{n-1})}{\tau} - \partial_t u(t_n) = -\frac{1}{\tau}\int_{J_n}(t - t_{n-1})\partial_{tt}u(t)\mathrm{d}t,$$

where the last equality follows by integrating by parts in time. The above equalities are meaningful in V' owing to the smoothness assumptions on u. Notice in particular that $H^2(J;V') \hookrightarrow C^1(\overline{J};V')$. Subtracting the above identity from (67.3) and using the definition of e_h^n, we have for all $w_h \in V_h$,

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau a^n(e_h^n, w_h) = \tau\langle g^n, w_h\rangle_{V',V},$$

where

$$\begin{aligned}\langle g^n, w_h\rangle_{V',V} &:= a^n(\eta(t_n), w_h) + \tau^{-1}(\eta(t_n) - \eta(t_{n-1}), w_h)_L - \langle \psi^n, w_h\rangle_{V',V}\\ &= a^n(\eta(t_n), w_h) + \langle \xi^n - \psi^n, w_h\rangle_{V',V},\end{aligned}$$

with $\xi^n := \frac{1}{\tau}\int_{J_n} \partial_t\eta(t)\mathrm{d}t$.
(3) Applying the stability estimate from Lemma 67.3 with e_h^n in lieu of u_h^n and g^n in lieu of f^n and dropping the terms related to the time increment and the value of e_h at the final time for simplicity, we infer that

$$\alpha\|e_{h\tau}\|_{\ell^2(J;V)}^2 \le \frac{1}{\alpha}\|g_\tau\|_{\ell^2(J;V')}^2 + \|e_h^0\|_L^2, \tag{67.13}$$

with $e_{h\tau} := (e_h^n)_{n\in\mathcal{N}_\tau}$ and $g_\tau := (g^n)_{n\in\mathcal{N}_\tau}$. The boundedness of a^n, the triangle inequality, and the Cauchy–Schwarz inequality imply that

$$\begin{aligned}\|g_\tau\|_{\ell^2(J;V')}^2 &= \sum_{n\in\mathcal{N}_\tau} \tau\|g^n\|_{V'}^2\\ &\le \sum_{n\in\mathcal{N}_\tau} \tau(M\|\eta(t_n)\|_V + \tau^{-\frac12}\|\partial_t\eta\|_{L^2(J_n;V')} + \tau^{\frac12}\|\partial_{tt}u\|_{L^2(J_n;V')})^2\\ &\le 3\big(M^2\|\eta_\tau\|_{\ell^2(J;V)}^2 + \|\partial_t\eta\|_{L^2(J;V')}^2 + \tau^2\|\partial_{tt}u\|_{L^2(J;V')}^2\big).\end{aligned}$$

(4) Taking the square root of (67.13), using the triangle inequality on the error $u_\tau - u_{h\tau} = \eta_\tau - e_{h\tau}$, and using that $\|e_h^0\|_L \le \|\eta(0)\|_L$ since $u_h^0 := \mathcal{P}_{V_h}(u_0)$ readily yields the assertion. □

Remark 67.7 (Estimate (67.12)). If $u \in H^1(J;V)\cap H^2(J;L)$, then using $\|\phi\|_{V'} \le \iota_{L,V}\|\phi\|_L$ for all $\phi \in L$, and the definition of the time scale ρ, (67.12) implies that

$$\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le c_1'\tau\|\partial_{tt}u\|_{L^2(J;L)} + c_2\|\eta_\tau\|_{\ell^2(J;V)} + c_1'\|\partial_t\eta\|_{L^2(J;L)} + c_3\|\eta(0)\|_L, \tag{67.14}$$

with $c_1' := \sqrt{\frac{3\rho}{2\alpha}}$, and c_2, c_3 as in (67.12). The first term on the right-hand side of (67.12) and (67.14) is related to the discretization error in time and converges as $\mathcal{O}(\tau)$, i.e., the method is first-order accurate in time. The other three terms, which involve the function η, are related to the discretization error in space measured in various norms. Notice also that owing to the bound $\|v\|_{L^2(J;B)} \le \sqrt{T}\|v\|_{L^\infty(J;B)}$, (67.12) implies that

$$\begin{aligned}\frac{1}{\sqrt{T}}\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} &\le c_1\tau\|\partial_{tt}u\|_{L^\infty(J;V')}\\ &\quad + \frac{c_2}{\sqrt{T}}\|\eta_\tau\|_{\ell^2(J;V)} + c_1\|\partial_t\eta\|_{L^\infty(J;V')} + \frac{c_3}{\sqrt{T}}\|\eta(0)\|_L,\end{aligned}$$

under the slightly stronger assumption $u \in H^1(J;V)\cap W^{2,\infty}(J;V')$. A similar variant can be established for (67.14). □

Remark 67.8 (Supercloseness). Assuming for simplicity that the bilinear form a is time-independent, another interesting choice for the error decomposition is to set $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$, where $\Pi_h^{\mathrm{E}} : V \to V_h$ is the elliptic projection defined in (66.14) (that is, $a(\Pi_h^{\mathrm{E}}(v), w_h) := a(v, w_h)$ for all $v \in V$ and all $w_h \in V_h$). This gives the coefficients $c_1 := \frac{\sqrt{2}}{\alpha}$, $c_2 := 1$, and $c_3 := \frac{1}{\sqrt{\alpha}}$ in (67.12). More importantly, we obtain a supercloseness estimate on the discrete error (see Remark 66.8), i.e., setting $\Pi_h^{\mathrm{E}}(u)_\tau := (\Pi_h^{\mathrm{E}}(u(t_n)))_{n\in\mathcal{N}_\tau}$, we have

$$\|\Pi_h^{\mathrm{E}}(u)_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le c_1\tau\|\partial_{tt}u\|_{L^2(J;V')} + c_1\|\partial_t\eta\|_{L^2(J;V')} + c_3\|\eta(0)\|_L,$$

which avoids the $\ell^2(J;V)$-norm in the upper bound. □

We now consider the error estimates in the $\ell^\infty(\overline{J};L)$-norm. For simplicity, we invoke the embedding $L \hookrightarrow V'$, and we use the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$.

Theorem 67.9 (Improved $\ell^\infty(\overline{J};L)$-estimate). *Let u solve the parabolic problem (65.10) and let $u_{h\tau} \in (V_h)^N$ solve (67.3). Assume that $u \in C^0(\overline{J};V) \cap H^2(J;L)$. Assume that $\tau \le \frac{1}{2}\rho$ (for simplicity) and that the bilinear form a is time-independent. The following holds true for all $n \in \mathcal{N}_\tau$ and all $h \in \mathcal{H}$:*

$$\begin{aligned}\|u(t_n) - u_h^n\|_L \le \|\eta(t_n)\|_L + e^{-\frac{t_n}{2\rho}}\|\eta(0)\|_L \\ + \sqrt{\rho}\big(\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t\eta\|_{L^2((0,t_n);L)} + \tau\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_{tt}u\|_{L^2((0,t_n);L)}\big),\end{aligned} \tag{67.15}$$

where $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$ and $\Pi_h^{\mathrm{E}} : V \to V_h$ is the elliptic projection defined in (66.14).

Proof. We start as in the proof of Theorem 67.6, and we make the choice $v_h(t) := \Pi_h^{\mathrm{E}}(u(t))$. The crucial point is that $a(\eta(t_n), w_h) = 0$ for all $n \in \mathcal{N}_\tau$ and all $w_h \in V_h$. We infer that

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau a^n(e_h^n, w_h) = \tau\langle \xi^n - \psi^n, w_h\rangle_{V',V},$$

where we recall that $\xi^n := \frac{1}{\tau}\int_{J_n}\partial_t\eta(t)\mathrm{d}t$ and $\psi^n := -\frac{1}{\tau}\int_{J_n}(t-t_{n-1})\partial_{tt}u(t)\mathrm{d}t$. We now invoke Lemma 67.5 and infer that for all $n \in \mathcal{N}_\tau$,

$$\|e_h^n\|_L^2 \le e^{-\frac{t_n}{\rho}}\|e_h^0\|_L^2 + \rho\sum_{k\in\{1:n\}} e^{-\frac{t_n-t_{k-1}}{\rho}}\big(\|\partial_t\eta\|_{L^2(J_k;L)}^2 + \tau^2\|\partial_{tt}u\|_{L^2(J_k;L)}^2\big),$$

where we used that

$$\frac{1}{\alpha}\tau\|\xi^k - \psi^k\|_{V'}^2 \le \frac{\tau\rho}{2}\|\xi^k - \psi^k\|_L^2 \le \rho\big(\|\partial_t\eta\|_{L^2(J_k;L)}^2 + \tau^2\|\partial_{tt}u\|_{L^2(J_k;L)}^2\big).$$

Since $e^{-\frac{t_n-t_{k-1}}{\rho}} \le e^{-\frac{t_n-s}{\rho}}$ for all $s \in J_k$, we obtain

$$\|e_h^n\|_L^2 \le e^{-\frac{t_n}{\rho}}\|e_h^0\|_L^2 + \rho\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t\eta\|_{L^2((0,t_n);L)}^2 + \tau^2\rho\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_{tt}u\|_{L^2((0,t_n);L)}^2.$$

Taking the square root of this estimate, recalling that $\|e_h^0\|_L \le \|\eta(0)\|_L$, and invoking the triangle inequality for the error $u(t_n) - u_h^n = \eta(t_n) - e_h^n$ proves the assertion. □

Remark 67.10 (Comparison). The improvement with respect to Theorem 67.6 is twofold. On the one hand we capture the exponential decay of the influence of the error induced by the approximation of the initial data. On the other hand the use of the elliptic projection removes the suboptimal term $\|\eta\|_{\ell^2((0,t_n);V)}$ from the error estimate. □

67.1.5 Application to the heat equation

Let us now particularize the setting to the heat equation with $V := H_0^1(D)$, $L := L^2(D)$, $V' = H^{-1}(D)$, $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$, and $\|v\|_L := \|v\|_{L^2(D)}$, so that $\iota_{L,V} = C_{\mathrm{PS}}^{-1}\ell_D$, where C_{PS} is the Poincaré–Steklov constant in $H_0^1(D)$ and ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$. The time scale becomes $\rho := \frac{2}{C_{\mathrm{PS}}^2}\frac{\ell_D^2}{\alpha}$.

Corollary 67.11 (Convergence rates). *Let $r \in [1,k]$, where $k \geq 1$ is the degree of the finite elements used to build the discrete space V_h.* (i) *Assume that $u \in C^0(\overline{J};H^{r+1}(D))\cap H^1(J;H^r(D))\cap H^2(J;L^2(D))$. There are c_1, c_2 such that for all $h\in\mathcal{H}$, τ, T, α, and M,*

$$\|u_\tau - u_{h\tau}\|_{\ell^2(J;H_0^1(D))} \leq c_1\,\tau\frac{\rho}{\ell_D}\|\partial_{tt}u\|_{L^2(J;L^2(D))} + c_2\,h^r\Big(\frac{M}{\alpha}|u_\tau|_{\ell^2(J;H^{r+1}(D))} + \frac{\rho}{\ell_D}|\partial_t u|_{L^2(J;H^r(D))} + \frac{1}{\sqrt{\alpha}}|u_0|_{H^r(D)}\Big). \tag{67.16}$$

(ii) *Assume that there is some elliptic regularity pickup in the associated adjoint problem, i.e., there are $s\in(0,1]$ and $c_{\rm smo}>0$ such that for all $g\in L^2(D)$, the unique function $\xi_g\in H_0^1(D)$ s.t. $a(v,\xi_g)=(g,v)_{L^2(D)}$ for all $v\in H_0^1(D)$ satisfies $\|\xi_g\|_{H^{1+s}(D)}\leq c_{\rm smo}\alpha^{-1}\ell_D^2\|g\|_{L^2(D)}$. Assume that $u\in H^1(J;H^{r+1}(D))\cap H^2(J;L^2(D))$ and (for simplicity) $\tau\leq\frac12\rho$. Then there is c such that for all $h\in\mathcal{H}$, τ, T, α, M, and all $n\in\mathcal{N}_\tau$,*

$$\begin{aligned}\|u(t_n)-u_h^n\|_{L^2(D)} \leq{}& \tau\sqrt{\rho}\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_{tt}u\|_{L^2((0,t_n);L^2(D))}\\ &+ c\,h^{r+s}\ell_D^{1-s}\Big(\frac{M}{\alpha}\Big)^2\Big(|u(t_n)|_{H^{r+1}(D)} + e^{-\frac{t_n}{2\rho}}|u_0|_{H^{r+1}(D)}\\ &+\sqrt{\rho}|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t u|_{L^2((0,t_n);H^{r+1}(D))}\Big).\end{aligned} \tag{67.17}$$

Proof. To prove (67.16), we start from (67.14) and proceed as in Corollary 66.12 to estimate the terms involving η. To prove (67.17), we use Theorem 67.9 and proceed as in Corollary 66.13 by using the approximation properties of the elliptic projection. □

Remark 67.12 (Corollary 67.11). The estimate (67.16) exhibits the optimal decay rate $\mathcal{O}(h^r+\tau)$. Notice that we are using the seminorms $|u_\tau|_{\ell^2(J;H^{r+1}(D))}$ and $|\partial_t u|_{L^2(J;H^r(D))}$. Moreover, it is possible to localize the right-hand side of (67.16) to the mesh cells, and it is also possible to make the weaker smoothness requirement $u\in C^0(\overline{J};H^{r+1}(D))\cap H^1(J;H^{r-1}(D))\cap H^2(J;H^{-1}(D))$ by starting from (67.12) instead of (67.14) (i.e., avoiding the embedding $L^2(D)\hookrightarrow H^{-1}(D)$). Furthermore, the estimate (67.17) exhibits the quasi-optimal decay rate $\mathcal{O}(h^{r+s}+\tau)$, and this rate is optimal if there is full elliptic regularity pickup, i.e., the rate is $\mathcal{O}(h^{r+1}+\tau)$ if $s=1$. Notice that the seminorm $|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t u|_{L^2((0,t_n);H^{r+1}(D))}$ is used in (67.17). As in the semi-discrete setting of §66.3.1, the use of the elliptic projection is crucial to achieve quasi-optimal decay rates. Finally, we observe that all the terms on the right-hand side of (67.17) (excluding the factor h^s) can be localized to the mesh cells. □

67.2 Explicit Euler scheme

In this section, we briefly discuss the *explicit Euler scheme*. For brevity, we focus on the main stability and error estimates. The salient difference with the implicit Euler scheme is that the linear algebra involved at each time step is simpler since the inversion of a matrix involving the stiffness matrix is avoided. But this gain in simplicity is traded against stability since now the scheme becomes *conditionally stable*, that is, stability requires that the time step be smaller than a constant times some power of the mesh size. In the context of the heat equation, the upper bound on the time step scales as the square of the meshsize (for quasi-uniform mesh sequences).

67.2.1 Principle and algebraic realization

We use the notation from §67.1.1 for the time discretization, and the notation from §67.1.2 for the space discretization. As above, the space discretization is done using a sequence of finite-dimensional and time-independent spaces $(V_h)_{h\in\mathcal{H}}$, but to discretize the time derivative we now write $\partial_t u_h(t_{n-1}) = \frac{u_h(t_n)-u_h(t_{n-1})}{\tau} + \mathcal{O}(\tau)$ for all $n \in \mathcal{N}_\tau$. After setting $u_h^0 := u_h(0) = \mathcal{P}_{V_h}(u_0)$, as for the implicit Euler scheme, the discrete problem consists of seeking a sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ s.t. for all $n \in \mathcal{N}_\tau$,

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a^{n-1}(u_h^{n-1}, w_h) = \tau\langle f^{n-1}, w_h\rangle_{V',V},\ \forall w_h \in V_h, \tag{67.18}$$

where $a^{n-1}(\cdot,\cdot) := a(t_{n-1};\cdot,\cdot)$ and $f^{n-1} := f(t_{n-1}) \in V'$.

Let U^n, U^{n-1} be the coordinate vectors of u_h^n and u_h^{n-1} in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, respectively. Recalling the stiffness matrix $\mathcal{A}(t) \in \mathbb{R}^{I\times I}$ and the mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$ defined in (67.4), the explicit Euler scheme (67.18) is recast as follows: For all $n \in \mathcal{N}_\tau$,

$$\mathcal{M}(\mathsf{U}^n - \mathsf{U}^{n-1}) + \tau\mathcal{A}^{n-1}\mathsf{U}^{n-1} = \tau\mathsf{F}^{n-1}, \tag{67.19}$$

with $\mathcal{A}^{n-1} := \mathcal{A}(t_{n-1})$ and $\mathsf{F}^{n-1} := (\langle f^{n-1}, \varphi_i\rangle_{V',V})_{i\in\{1:I\}}$. Rearranging the terms in (67.19) leads to

$$\mathcal{M}\mathsf{U}^n = (\mathcal{M} - \tau\mathcal{A}^{n-1})\mathsf{U}^{n-1} + \tau\mathsf{F}^{n-1}, \tag{67.20}$$

showing that each step of the explicit Euler scheme entails solving a linear system associated with the symmetric positive definite mass matrix $\mathcal{M}$. Inverting the mass matrix is significantly easier than solving a linear system involving the stiffness matrix. Indeed, $\mathcal{M}$ is always symmetric and has better conditioning properties than $\mathcal{M}+\tau\mathcal{A}^n$ (see §28.2.1). Notice also that $\mathcal{M}$ is time-independent.

67.2.2 Stability

The stability analysis of the explicit Euler time-stepping method depends on the following mesh-dependent parameter:

$$c_{\mathrm{INV}}(h) := \iota_{L,V} \max_{v_h\in V_h} \frac{\|v_h\|_V}{\|v_h\|_L}. \tag{67.21}$$

This quantity is nondimensional and it is finite since V_h is finite-dimensional. For the heat equation, we have $V := H_0^1(D)$, $L := L^2(D)$, with $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$, $\|v\|_L := \|v\|_{L^2(D)}$, so that $\iota_{L,V} := C_{\mathrm{PS}}^{-1}\ell_D$ where C_{PS} is the Poincaré–Steklov constant in $H_0^1(D)$ and ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. If V_h is a finite element space based on a quasi-uniform mesh sequence, the inverse inequality in Lemma 12.1 shows that $c_{\mathrm{INV}}(h) \le c\ell_D h^{-1}$ for all $h \in \mathcal{H}$. On a shape-regular mesh sequence, the constant $c_{\mathrm{INV}}(h)$ scales like the inverse of the diameter of the smallest mesh cell.

Recall that α and M denote the coercivity and the boundedness constants of the bilinear form a. We use the notation $v_{h\tau} := (v_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$, $v_{h\tau}^- := (v_h^{n-1})_{n\in\mathcal{N}_\tau} \in (V_h)^N$, and $\delta_\tau v_{h\tau} := (\frac{1}{\tau}(v_h^n - v_h^{n-1}))_{n\in\mathcal{N}_\tau} \in (V_h)^N$, and we use the time-discrete norms defined in (67.1). We consider as above the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$.

Lemma 67.13 ($\ell^2(J;V)$- and $\ell^\infty(\overline{J};L)$-stability). *Let the sequence $u_{h\tau} \in (V_h)^N$ solve (67.18) with the sequence of source terms $f_\tau^- := (f^{n-1})_{n\in\mathcal{N}_\tau} \in (V')^N$. Let $\xi_\kappa := \frac{M}{\alpha}$.* (i) *Assume that τ is small enough so that the following parabolic CFL condition holds true:*

$$\tau \le \frac{1}{8}\rho\xi_\kappa^{-2}c_{\mathrm{INV}}(h)^{-2}. \tag{67.22}$$

The following stability estimate holds true:

$$\alpha\|u_{h\tau}\|^2_{\ell^2(J;V)} + \frac{\tau}{2}\|\delta_\tau u_{h\tau}\|^2_{\ell^2(J;L)} + \|u_h^N\|_L^2 \le \frac{2}{\alpha}\|f_\tau^-\|^2_{\ell^2(J;V')} + \|u_h^0\|_L^2. \tag{67.23}$$

(ii) *If in addition to* (67.22)*, the time step also satisfies* $\tau \le \frac{1}{2}\rho$*, we have for all* $n \in \mathcal{N}_\tau$*,*

$$\|u_h^n\|_L^2 \le e^{-\frac{t_n}{\rho}}\|u_h^0\|_L^2 + \frac{2}{\alpha}\sum_{k\in\{1:n\}} \tau e^{-\frac{t_n - t_{k-1}}{\rho}}\|f^{k-1}\|^2_{V'}. \tag{67.24}$$

Proof. (1) Using $w_h := u_h^n$ as the test function in (67.18) and using the identity (67.9) leads to

$$\frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 + \frac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2 + \tau a^{n-1}(u_h^{n-1}, u_h^n) = \tau\langle f^{n-1}, u_h^n\rangle_{V',V}.$$

Using Young's inequality on the right-hand side leads to

$$\frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 + \frac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2 + \tau a^{n-1}(u_h^{n-1}, u_h^n) - \frac{1}{4}\alpha\tau\|u_h^n\|_V^2 \le \frac{1}{\alpha}\tau\|f^{n-1}\|^2_{V'}. \tag{67.25}$$

We observe that

$$\begin{aligned} a^{n-1}(u_h^{n-1}, u_h^n) &= a^{n-1}(u_h^n, u_h^n) - a^{n-1}(u_h^n - u_h^{n-1}, u_h^n) \\ &\ge \alpha\,\|u_h^n\|_V^2 - M\,\|u_h^n - u_h^{n-1}\|_V\|u_h^n\|_V \\ &\ge \alpha\,\|u_h^n\|_V^2 - M\iota_{L,V}^{-1}c_{\mathrm{INV}}(h)\,\|u_h^n - u_h^{n-1}\|_L\|u_h^n\|_V \\ &\ge \frac{3}{4}\alpha\,\|u_h^n\|_V^2 - \frac{M^2}{\alpha}\iota_{L,V}^{-2}c_{\mathrm{INV}}(h)^2\,\|u_h^n - u_h^{n-1}\|_L^2, \end{aligned}$$

where we used the coercivity and the boundedness of a in the second line, the inverse inequality (67.21) in the third line, and Young's inequality in the fourth line. Inserting this lower bound in (67.25) and since $\frac{M^2}{\alpha}\iota_{L,V}^{-2} = 2\xi_\kappa^2\rho^{-1}$ with $\xi_\kappa := \frac{M}{\alpha}$ and $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, we infer that

$$\frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 + \left(\frac{1}{2} - 2\tau\rho^{-1}\xi_\kappa^2 c_{\mathrm{INV}}(h)^2\right)\|u_h^n - u_h^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\,\|u_h^n\|_V^2 \le \frac{1}{\alpha}\tau\|f^{n-1}\|^2_{V'}. \tag{67.26}$$

Using the bound (67.22) on τ and summing over $n \in \mathcal{N}_\tau$ proves (67.23).
(2) The proof of (67.24) starts from (67.26) and proceeds by using the same arguments as in the proof of Lemma 67.5. □

Remark 67.14 (Parabolic CFL). In the context of the heat equation, the upper bound on the time step resulting from (67.22) scales as $h^2\ell_D^{-2}$. In practice, this is often a quite restrictive limitation on the time step. Notice also that the upper bound in (67.22) depends on the model parameters α and M and becomes more stringent if the diffusion coefficient is highly contrasted (recall that α and M are global lower and upper bounds on this coefficient). This bound also scales as k^{-4}, where k is the polynomial degree of the finite elements (see the discussion at the end of §12.1). Notice also that the factor $\frac{1}{8}$ in (67.22) can be replaced by $\frac{1}{4}$ if one completely discards the nonessential term $\frac{\tau}{2}\|\delta_\tau u_{h\tau}\|^2_{\ell^2(J;L)}$ from the left-hand side of (67.23). □

67.2.3 Error analysis

The error analysis for the explicit Euler scheme is similar to that of the implicit Euler scheme. To avoid the proliferation of estimates, we just consider the error bounds with L-valued time derivatives by invoking the embedding $L \hookrightarrow V'$. Error bounds with V'-valued time derivatives can also be derived by proceeding as in Theorem 67.6. Recall that $\mathcal{I}_h : V \to V_h$ is a generic operator having optimal approximation properties.

Theorem 67.15 ($\ell^2(J;V)$- and $\ell^\infty(\overline{J};L)$-estimates). *Let u solve the model problem (65.10) and assume that $u \in H^1(J;V) \cap H^2(J;L)$. Let $\eta(t) := u(t) - \mathcal{I}_h(u(t))$ for all $t \in J$. Let us set $u_\tau := (u(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$ and $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$. Let $u_{h\tau} \in (V_h)^N$ solve (67.18). (i) Assume that the time step τ satisfies the parabolic CFL condition (67.22). Then we have*

$$\begin{aligned}\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le{}& c_1\tau\|\partial_{tt}u\|_{L^2(J;L)}\\ &+ \|\eta_\tau\|_{\ell^2(J;V)} + c_2\|\eta_\tau^-\|_{\ell^2(J;V)} + c_1\|\partial_t\eta\|_{L^2(J;L)} + c_3\|\eta(0)\|_L,\end{aligned}$$

with $c_1 := \frac{\sqrt{3\rho}}{\sqrt{\alpha}}$, $c_2 := \frac{\sqrt{6}M}{\alpha}$, and $c_3 := \frac{1}{\sqrt{\alpha}}$ with $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$. (ii) If in addition to (67.22) the time step also satisfies $\tau \le \frac{1}{2}\rho$ and if the bilinear form a is time-independent, then letting $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(u(t))$, where $\Pi_h^{\mathrm{E}} : V \to V_h$ is the elliptic projection defined in (66.14), we have for all $n \in \mathcal{N}_\tau$,

$$\begin{aligned}\|u(t_n) - u_h^n\|_L \le{}& \tau\sqrt{2\rho}\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_{tt}u\|_{L^2((0,t_n);L)}\\ &+ \|\eta(t_n)\|_L + e^{-\frac{t_n}{2\rho}}\|\eta(0)\|_L + \sqrt{2\rho}\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t\eta\|_{L^2((0,t_n);L)}.\end{aligned}$$

Proof. (1) We follow the proof of Theorem 67.6. Let us set $\eta(t) := u(t) - v_h(t)$ for all $t \in \overline{J}$, and $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau}$, where v_h is arbitrary in $H^1(J;V_h)$. For all $n \in \mathcal{N}_\tau$, we set $e_h^n := u_h^n - v_h(t_n)$ and obtain the error decomposition $u(t_n) - u_h^n = \eta(t_n) - e_h^n$. A straightforward calculation shows that for all $w_h \in V_h$,

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau a^{n-1}(e_h^{n-1}, w_h) = \tau\langle g^{n-1}, w_h\rangle_{V',V},$$

with $\langle g^{n-1}, w_h\rangle_{V',V} := a^{n-1}(\eta(t_{n-1}), w_h) + \langle \xi^{n-1} - \psi^{n-1}, w_h\rangle_{V',V}$, $\xi^{n-1} := \frac{1}{\tau}\int_{J_n}\partial_t\eta(t)\mathrm{d}t$, and $\psi^{n-1} := \frac{1}{\tau}\int_{J_n}(t_n - t)\partial_{tt}u(t)\mathrm{d}t$. Invoking now the stability estimate (67.23), we infer that

$$\alpha\|e_{h\tau}\|_{\ell^2(J;V)}^2 \le \frac{2}{\alpha}\|g_\tau^-\|_{\ell^2(J;V')}^2 + \|e_h^0\|_L^2.$$

The error estimate on $\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)}$ follows from the same arguments as in the proof of Theorem 67.6.
(2) The proof of the estimate on $\|u(t_n) - u_h^n\|_L$ is similar to that of Theorem 67.9, except that we now invoke the stability estimate (67.24). □

Example 67.16 (Heat equation). Assume that we are approximating the heat equation with H^1-conforming finite elements and with the explicit Euler scheme under the parabolic CFL restriction (67.22). Then the bounds from Theorem 67.15 imply that the error estimates (67.16) and (67.17) still hold true, that is, the error in the $\ell^2(J;H_0^1(D))$-norm decays as $\mathcal{O}(h^r + \tau)$ and the error in the $\ell^\infty(\overline{J};L^2(D))$-norm decays as $\mathcal{O}(h^{r+s} + \tau)$, where $s \in (0,1]$ is the elliptic regularity pickup index ($s = 1$ if there is full elliptic regularity pickup). □

Exercises

Exercise 67.1 (Incremental Gronwall's lemma). Let $\gamma \in \mathbb{R}$, $\gamma > -1$. Let $(a_n)_{n\in\overline{\mathcal{N}}_\tau}$, $(b_n)_{n\in\mathcal{N}_\tau}$ be two sequences of real numbers s.t. $(1+\gamma)a_n \le a_{n-1} + b_n$ for all $n \in \mathcal{N}_\tau$. Prove that $a_n \le \frac{a_0}{(1+\gamma)^n} + \sum_{k\in\{1:n\}} \frac{b_k}{(1+\gamma)^{n-k+1}}$ for all $n \in \mathcal{N}_\tau$. (*Hint*: by induction.) *Note*: it is common to use the above estimate together with the inequality $\frac{1}{1+\gamma} \le e^{-\frac{\gamma}{2}}$ for $\gamma \in (0,1)$. The reader is referred to Exercise 68.3 for a discrete form of the Gronwall using an assumption that is weaker than requesting that $(1+\gamma)a_n \le a_{n-1} + b_n$.

Exercise 67.2 (Inf-sup condition). Let $X_{h\tau} := (V_h)^{N+1}$ and $Y_{h\tau} := V_h \times (V_h)^N$. Define $\|\phi_h\|_{V_h'} := \sup_{v_h\in V_h} \frac{|(\phi_h, v_h)_L|}{\|v_h\|_V}$ for all $\phi_h \in V_h$ and consider the following norms:

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \frac{1}{\alpha}\|v_h^N\|_L^2 + \|v_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{\alpha M}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;V_h')}^2 + \frac{\tau}{\alpha}\|\delta_\tau v_{h\tau}\|_{\ell^2(J;L)}^2,$$
$$\|y_{h\tau}\|_{Y_{h\tau}}^2 := \frac{1}{\alpha}\|y_{0h}\|_L^2 + \|y_{1h\tau}\|_{\ell^2(J;V)}^2,$$

with $(\delta_\tau v_{h\tau})^n := \frac{1}{\tau}(v_h^n - v_h^{n-1})$, for all $v_{h\tau} \in X_{h\tau}$ and all $y_{h\tau} := (y_{0h}, y_{1h\tau}) \in Y_{h\tau}$. Define the bilinear form $b_\tau : X_{h\tau}\times Y_{h\tau} \to \mathbb{R}$ s.t.

$$b_\tau(v_{h\tau}, y_{h\tau}) := (v_h^0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\big(((\delta_\tau v_{h\tau})^n, y_{1h}^n)_L + a^n(v_h^n, y_{1h}^n)\big).$$

Assume that a is symmetric. The goal is to prove the following inf-sup condition:

$$\inf_{v_{h\tau}\in X_{h\tau}} \sup_{y_h\in Y_{h\tau}} \frac{|b_\tau(v_{h\tau}, y_{h\tau})|}{\|v_{h\tau}\|_{X_{h\tau}}\|y_{h\tau}\|_{Y_{h\tau}}} \ge \alpha\left(\frac{\alpha}{M}\right)^{\frac{1}{2}}. \tag{67.27}$$

(i) Let $A_h^n : V_h \to V_h'$ be s.t. $\langle A_h^n(z_h), w_h\rangle_{V_h',V_h} := a^n(z_h, w_h)$ for all $z_h, w_h \in V_h$ and all $n \in \mathcal{N}_\tau$. Consider the test function $w_{h\tau} := (w_{0h}, w_{1h\tau}) \in Y_{h\tau}$ with $w_{0h} := v_h^0$ and $w_{1h}^n := (A_h^n)^{-1}((\delta_\tau v_{h\tau})^n) + v_h^n$ for all $n \in \mathcal{N}_\tau$. Prove that $b_\tau(v_{h\tau}, w_{h\tau}) \ge \alpha\|v_{h\tau}\|_{X_{h\tau}}^2$. (*Hint*: use that $(A_h^n)^{-1}$ is coercive on V_h' with constant M^{-1}, see Lemma C.63.) (ii) Prove that $\alpha\tau\|w_{1h}^n\|_V^2 \le M\tau\|v_h^n\|_V^2 + \frac{\tau}{\alpha}\|(\delta_\tau v_{h\tau})^n\|_{V_h'}^2 + \|v_h^n\|_L^2 - \|v_h^{n-1}\|_L^2 + \tau^2\|(\delta_\tau v_{h\tau})^n\|_L^2$. (*Hint*: use the boundedness of $(A_h^n)^{-1}$ on V_h' with constant α^{-1}.) (iii) Conclude. *Note*: let $\mathfrak{T}_1 := \tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2$ and consider the bound on $\mathfrak{T}_1$ given in Lemma 67.3. Let $\mathfrak{T}_2 := \frac{1}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;V_h')}^2$ and consider the bound on $\mathfrak{T}_2$ given by the inf-sup condition (67.27) (see Exercise 71.8). If the functions $(\partial_t u(t_n))_{n\in\mathcal{N}_\tau}$ are smooth in space for all $n \in \mathcal{N}_\tau$, one expects that $\mathfrak{T}_2 \approx \frac{\iota_{L,V}^2}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2 = \frac{\rho}{2\tau}\frac{\alpha}{M}\mathfrak{T}_1$ with the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$. Hence, $\mathfrak{T}_2 \gg \mathfrak{T}_1$ if $\rho \gg \tau$, i.e., controlling $\mathfrak{T}_2$ is more informative than just controlling $\mathfrak{T}_1$.

Exercise 67.3 (Implicit-explicit scheme). Let $(V, L \equiv L', V')$ be a Gelfand triple. Let $B \in \mathcal{L}(V;L)$ and $A \in \mathcal{L}(V;V')$ be two operators. Assume that A is V-coercive with $\langle A(v), v\rangle_{V',V} \ge \alpha\|v\|_V^2$ for all $v \in V$, and that $\|v\|_L \le \iota_{L,V}\|v\|_V$. Let $\mathfrak{c}$ be s.t. $\mathfrak{c} \ge \max(\|B\|_{\mathcal{L}(V;L)}, \|B^*\|_{\mathcal{L}(L;V')})$. Let $u_0 \in V$ and $f \in C^0(\overline{J}; V')$. Consider the model problem $\partial_t u(t) + A(u)(t) + B(u)(t) = f(t)$ in $L^2(J;V')$, and $u(0) = u_0$. (i) Let $\nu > 0$, $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D)$, $u_0 \in L^2(D)$, and $f \in C^0(\overline{J}; H^{-1}(D))$. Show that the time-dependent advection-diffusion equation $\partial_t u - \nu\Delta u + \boldsymbol{\beta}\cdot\nabla u = f$, $u_{|\partial D} = 0$, $u(0) = u_0$ fits the above setting, i.e., specify the spaces V, L, the operators A, B, and the constants α, $\mathfrak{c}$ in this case. (ii) Let $f^n := f(t_n)$ for all $n \in \mathcal{N}_\tau$. Consider the following scheme: $u^0 := u_0$ and for all $v \in V$ and all $n \in \mathcal{N}_\tau$,

$$(u^n - u^{n-1}, v)_L + \tau\langle A(u^n), v\rangle_{V',V} + \tau(B(u^{n-1}), v)_L = \tau\langle f^n, v\rangle_{V',V}.$$

Prove that if $2\frac{\mathfrak{c}\iota_{L,V}}{\alpha} \leq 1$, then

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \leq \|u^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\|u^{n-1}\|_V^2 + 2\frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

(iii) Assume that $(B(v), v)_L \geq 0$ for all $v \in V$, and that the time step satisfies the bound $\tau \leq \frac{1}{2}\frac{\alpha}{\mathfrak{c}^2}$. (We no longer assume that $2\frac{\mathfrak{c}\iota_{L,V}}{\alpha} \leq 1$.) Prove that

$$\|u^n\|_L^2 + \alpha\tau\|u^n\|_V^2 \leq \|u^{n-1}\|_L^2 + \frac{1}{2}\alpha\tau\|u^{n-1}\|_V^2 + \frac{\tau}{\alpha}\|f^n\|_{V'}^2.$$

Chapter 68

BDF2 and Crank–Nicolson schemes

In this chapter, we discuss two time-stepping techniques that deliver second-order accuracy in time and, like the implicit Euler method, are unconditionally stable. One technique is based on a second-order backward differentiation formula (BDF2), and the other, called Crank–Nicolson, is based on the midpoint quadrature rule. The BDF2 method is a *two-step* scheme, i.e., u_h^n is computed from u_h^{n-1} and u_h^{n-2} which are the approximations at the time nodes t_{n-1} and t_{n-2}. This feature makes the BDF2 method not well suited to time step adaptation. Moreover, the stability analysis must account on the way the scheme is initialized at the first time step (we use here an implicit Euler step). In contrast to this, the Crank–Nicolson scheme, like the implicit Euler scheme, is a *one-step* method, i.e., u_h^n only depends on the preceding time approximation u_h^{n-1} at the time node t_{n-1}. We will see however that the stability properties of the Crank–Nicolson method are not as strong as those of the implicit Euler method.

68.1 Discrete setting

We use the notation introduced in §67.1.1 for the time discretization, and we consider as in §67.1.2 the sequence of finite-dimensional and time-independent spaces $(V_h)_{h\in\mathcal{H}}$ for the space discretization. The operator $\mathcal{P}_{V_h} : L \to V_h$ is the L-orthogonal projection, i.e., for all $z \in L$, $\mathcal{P}_{V_h}(z)$ is the unique element in V_h s.t. $(z - \mathcal{P}_{V_h}(z), w_h)_L := 0$ for all $w_h \in V_h$. Letting $N > 0$ be a positive natural number, recall that we divide the time interval $J := (0, T)$ with $T > 0$ into N subintervals J_n for all $n \in \mathcal{N}_\tau := \{1{:}N\}$. For simplicity, we assume that all these intervals are of equal length, i.e., we define the *time step* to be $\tau := \frac{T}{N}$. Letting $t_n := n\tau$ be the *discrete time nodes* for all $n \in \overline{\mathcal{N}}_\tau := \{0{:}N\}$, we set $J_n := (t_{n-1}, t_n)$ for all $n \in \mathcal{N}_\tau$. For simplicity, we assume in the entire chapter that $f \in C^0(\overline{J}; V')$ and that the bilinear form $a(t; \cdot, \cdot)$ is well defined for all $t \in \overline{J}$.

68.2 BDF2 scheme

We review in this section the time-stepping technique based on the second-order *backward differentiation formula* (BDF2).

68.2.1 Principle and algebraic realization

The idea is to approximate the time derivative using the BDF2 formula

$$\partial_t u_h(t_n) = \frac{3u_h(t_n) - 4u_h(t_{n-1}) + u_h(t_{n-2})}{2\tau} + \mathcal{O}(\tau^2),$$

for all $n \in \mathcal{N}_\tau$, $n \geq 2$. After setting $u_h^0 := \mathcal{P}_{V_h}(u_0)$, as for the Euler schemes, we construct the sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ such that

$$(u_h^1 - u_h^0, w_h)_L + \tau a^1(u_h^1, w_h) = \tau\langle f^1, w_h\rangle_{V',V}, \tag{68.1a}$$

$$(\tfrac{3}{2}u_h^n - 2u_h^{n-1} + \tfrac{1}{2}u_h^{n-2}, w_h)_L + \tau a^n(u_h^n, w_h) = \tau\langle f^n, w_h\rangle_{V',V}, \tag{68.1b}$$

for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, $n \geq 2$, with $a^n(\cdot,\cdot) := a(t_n;\cdot,\cdot)$ and $f^n := f(t_n) \in V'$ for all $n \in \mathcal{N}_\tau$. Notice that an implicit Euler step is used at the first time step. Other choices are possible for the initialization, for instance, one could use a second-order single-step implicit scheme as the Crank–Nicolson scheme from §68.3.

Recall the stiffness matrix $\mathcal{A}(t) \in \mathbb{R}^{I\times I}$ and the mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$ defined in (67.4), i.e., $\mathcal{A}_{ij}(t) := a(t;\varphi_j,\varphi_i)$ and $\mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L$ for all $i,j \in \{1{:}I\}$, where $\{\varphi_i\}_{i\in\{1:I\}}$ is a basis of V_h with $I := \dim(V_h)$ ($\{\varphi_i\}_{i\in\{1:I\}}$ are usually the global shape functions in V_h). Then the algebraic form of the BDF2 scheme (68.1) is as follows:

$$(\mathcal{M} + \tau\mathcal{A}^1)\mathsf{U}^1 = \mathcal{M}\mathsf{U}^0 + \tau\mathsf{F}^1, \tag{68.2a}$$

$$(\tfrac{3}{2}\mathcal{M} + \tau\mathcal{A}^n)\mathsf{U}^n = \mathcal{M}(2\mathsf{U}^{n-1} - \tfrac{1}{2}\mathsf{U}^{n-2}) + \tau\mathsf{F}^n, \tag{68.2b}$$

for all $n \in \mathcal{N}_\tau$, $n \geq 2$ with $\mathcal{A}^n := \mathcal{A}(t_n)$ and $\mathsf{F}^n := (\langle f^n, \varphi_i\rangle_{V',V})_{i\in\{1:I\}}$, and U^n is the coordinate vector of u_h^n in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, i.e., $u_h^n := \sum_{i\in\{1:I\}} \mathsf{U}_i^n\varphi_i$. The formula (68.2b) shows that the computational cost of one step of the BDF2 scheme is comparable to that of the implicit Euler scheme.

68.2.2 Stability

To account for the structure of the initialization in (68.1a), we assume that f^1 can be decomposed into $f^1 := f^1_{V'} + f^1_L$ with $f^1_{V'} \in V'$ and $f^1_L \in L$. We define the time sequence $\tilde{f}_\tau := (\tilde{f}^n)_{n\in\mathcal{N}_\tau} \in (V')^N$ such that

$$\tilde{f}^1 := f^1_{V'}, \qquad \tilde{f}^n := f^n, \quad \forall n \in \mathcal{N}_\tau,\ n \geq 2. \tag{68.3}$$

The decomposition $f^1 = f^1_{V'} + f^1_L$ may not be unique. One only requires that such a decomposition exists with the quantities $\|f^1_{V'}\|_{V'}$ and $\|f^1_L\|_L$ being finite. The key idea is that we are going to derive a stability estimate where the term $\|f^1_L\|_L$ has an additional factor $\alpha^{\frac{1}{2}}\tau$ with respect to the term $\|f^1_{V'}\|_{V'}$ (recall that α denotes the coercivity constant of the bilinear form a.)

Let us first establish a stability estimate by means of a coercivity argument. Recall the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$. We consider the time-discrete norm $\|\phi_\tau\|^2_{\ell^2(J;B)} := \sum_{n\in\mathcal{N}_\tau} \tau\|\phi^n\|_B^2$ with $\phi_\tau := (\phi^n)_{n\in\mathcal{N}_\tau} \in B^N$, $B := V$ or $B := V'$ (see (67.1)).

Lemma 68.1 ($\ell^2(J;V)$-stability). *Let $u_{h\tau} \in (V_h)^N$ solve (68.1) with the sequence of source terms $f_\tau := (f^n)_{n\in\mathcal{N}_\tau} \in (V')^N$. Let $f^1 = f^1_{V'} + f^1_L$ and $\tilde f_\tau \in (V')^N$ be defined in (68.3). The following holds true:*

$$\alpha\|u_{h\tau}\|^2_{\ell^2(J;V)} + \|u_h^N\|_L^2 \le \frac{2}{\alpha}\|\tilde f_\tau\|^2_{\ell^2(J;V')} + 29\tau^2\|f_L^1\|_L^2 + \frac{5}{2}\|u_h^0\|_L^2. \tag{68.4}$$

Proof. The proof is similar to that that of Lemma 67.3. Let us first consider the case $n = 1$. Let $\lambda \in (0,2)$ and $\mu \in (0,1)$. Taking $w_h := 2u_h^1$ in (68.1a), using the key identity (67.9) for $n = 1$, and Young's inequality to bound the right-hand side, we infer that

$$\begin{aligned}\|u_h^1\|_L^2 - \|u_h^0\|_L^2 + \|u_h^1 - u_h^0\|_L^2 + 2\alpha\tau\|u_h^1\|_V^2 &= 2\tau\langle f^1_{V'}, u_h^1\rangle_{V',V} + 2\tau(f_L^1, u_h^1)_L\\ &\le \frac{\tau}{\lambda\alpha}\|f^1_{V'}\|^2_{V'} + \lambda\alpha\tau\|u_h^1\|_V^2 + \frac{\tau^2}{\mu}\|f_L^1\|_L^2 + \mu\|u_h^1\|_L^2.\end{aligned}$$

Rearranging the terms leads to

$$(1-\mu)\|u_h^1\|_L^2 + \|u_h^1 - u_h^0\|_L^2 + (2-\lambda)\alpha\tau\|u_h^1\|_V^2 \le \frac{\tau}{\lambda\alpha}\|f^1_{V'}\|^2_{V'} + \frac{\tau^2}{\mu}\|f_L^1\|_L^2 + \|u_h^0\|_L^2.$$

Since $\frac{2}{3}(1-\mu) \le 1$, we have

$$\begin{aligned}(1-\mu)\|u_h^1\|_L^2 + \|u_h^1 - u_h^0\|_L^2 &\ge \frac{1-\mu}{3}(3\|u_h^1\|_L^2 + 2\|u_h^1 - u_h^0\|_L^2)\\ &= \frac{1-\mu}{3}(\|u_h^1\|_L^2 + \|2u_h^1 - u_h^0\|_L^2 + \|u_h^0\|_L^2),\end{aligned}$$

where we used the identity $\|2u_h^1 - u_h^0\|_L^2 + \|u_h^0\|_L^2 = 2\|u_h^1\|_L^2 + 2\|u_h^1 - u_h^0\|_L^2$. Putting the above two bounds together and rearranging the terms yields

$$\begin{aligned}\|u_h^1\|_L^2 + \|2u_h^1 - u_h^0\|_L^2 + \frac{3(2-\lambda)}{1-\mu}\alpha\tau\|u_h^1\|_V^2&\\ \le \frac{3}{(1-\mu)\lambda}\frac{\tau}{\alpha}\|f^1_{V'}\|^2_{V'} + \frac{3}{(1-\mu)\mu}\tau^2\|f_L^1\|_L^2 + \frac{2+\mu}{1-\mu}\|u_h^0\|_L^2.&\end{aligned} \tag{68.5}$$

Now we choose λ and μ so that $\frac{3(2-\lambda)}{1-\mu} = 1$ and $\frac{3}{(1-\mu)\lambda} = 2$, i.e., $\lambda = 1 + \frac{\sqrt{2}}{2} \approx 1.707$ and $\mu = 3\frac{\sqrt{2}}{2} - 2 \approx 0.121$. This gives

$$\|u_h^1\|_L^2 + \|2u_h^1 - u_h^0\|_L^2 + \alpha\tau\|u_h^1\|_V^2 \le \frac{2}{\alpha}\tau\|f^1_{V'}\|^2_{V'} + 29\tau^2\|f_L^1\|_L^2 + \frac{5}{2}\|u_h^0\|_L^2, \tag{68.6}$$

since $\frac{3}{(1-\mu)\mu} \approx 28.14 \le 29$ and $\frac{2+\mu}{1-\mu} \approx 2.41 \le \frac{5}{2}$. Let us now consider the case $n \ge 2$. We make use of the following identity:

$$\begin{aligned}2(3u_h^n - 4u_h^{n-1} + u_h^{n-2}, u_h^n)_L &= \|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2\\ &\quad + \|2u_h^n - u_h^{n-1}\|_L^2 - \|2u_h^{n-1} - u_h^{n-2}\|_L^2 + \|u_h^n - 2u_h^{n-1} + u_h^{n-2}\|_L^2,\end{aligned}$$

so that, taking the test function $w_h := 4u_h^n$ in (68.1b), we infer that

$$\begin{aligned}\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \|2u_h^n - u_h^{n-1}\|_L^2 - \|2u_h^{n-1} - u_h^{n-2}\|_L^2&\\ + \|u_h^n - 2u_h^{n-1} + u_h^{n-2}\|_L^2 + 2\alpha\tau\|u_h^n\|_V^2 \le \frac{2}{\alpha}\tau\|f^n\|^2_{V'}.&\end{aligned} \tag{68.7}$$

Summing this bound over $n \in \mathcal{N}_\tau$, $n \geq 2$, adding the bound (68.6), canceling the telescoping terms, and dropping the positive terms $\|2u_h^N - u_h^{N-1}\|_L^2$ and $\sum_{n\geq 2}\|u_h^n - 2u_h^{n-1} + u_h^{n-2}\|_L^2$ on the left-hand side gives (68.4). (Notice also that we dropped the factor 2 in front of $\alpha\tau\|u_h^n\|_V^2$ in (68.7) since this factor is not present in (68.6).) □

We now derive a sharper stability estimate in the $\ell^\infty(\overline{J};L)$-norm that captures the exponentially decaying influence of the data on the solution.

Lemma 68.2 ($\ell^\infty(\overline{J};L)$-stability, exponential decay). *Let $u_{h\tau} \in (V_h)^N$ solve (68.1) with the sequence of source terms $f_\tau := (f^n)_{n\in\mathcal{N}_\tau} \in (V')^N$. Let $\tilde{f}_\tau \in (V')^N$ be defined in (68.3). Assume that $\tau \leq \frac{1}{3}\rho$. The following holds true for all $n \in \mathcal{N}_\tau$:*

$$\|u_h^n\|_L^2 \leq e^{-\frac{t_n}{4\rho}}\Big(91\tau^2\|f_L^1\|_L^2 + \frac{5}{2}\|u_h^0\|_L^2\Big) + \frac{2}{\alpha}\sum_{k\in\{1:n\}}\tau e^{-\frac{t_n - t_{k-1}}{4\rho}}\|\tilde{f}^k\|_{V'}^2. \tag{68.8}$$

Proof. Let us set $\gamma := \frac{2}{7}\frac{\tau}{\rho}$. We are going to show that, provided $\gamma \leq \frac{2}{21}$, i.e., $\frac{\tau}{\rho} \leq \frac{1}{3}$, we have

$$(1+\gamma)a^n \leq a^{n-1} + b^n, \qquad \forall n \in \mathcal{N}_\tau, \tag{68.9}$$

where $a^n := \|u_h^n\|_L^2 + \|2u_h^n - u_h^{n-1}\|_L^2 + \frac{\alpha\tau}{1+\gamma}\|u_h^n\|_V^2$, $b^n := \frac{2}{\alpha}\tau\|\tilde{f}^n\|_{V'}^2$ for all $n \in \mathcal{N}_\tau$, and $a^0 := 91\tau^2\|f_L^1\|_L^2 + \frac{5}{2}\|u_h^0\|_L^2$. Let us first consider the case $n = 1$. We multiply (68.5) by $1+\gamma$, and we define λ and μ so that $\frac{3(2-\lambda)}{1-\mu} = \frac{1}{1+\gamma}$ and $\frac{3}{(1-\mu)\lambda} = \frac{2}{1+\gamma}$, i.e., $\lambda := 1 + \frac{\sqrt{2}}{2}$, $\mu(\gamma) := 1 - \frac{3(1+\gamma)}{2\lambda}$. Notice that $\mu(\gamma)$ is a decreasing function of γ, and $\mu(\gamma) > 0$ for all $\gamma \in (0,\gamma_*)$ with $\gamma_* := \frac{\sqrt{2}-1}{3}$. Here, we have chosen $\gamma \leq \frac{2}{21} \leq \gamma_*$ to fix the ideas. This yields

$$(1+\gamma)a^1 \leq \frac{2\lambda}{\mu(\gamma)}\tau^2\|f_L^1\|_L^2 + \frac{2\lambda(2+\mu(\gamma))}{3}\|u_h^0\|_L^2 + b^1.$$

Since $\frac{2\lambda}{\mu(\gamma)}$ is an increasing function of γ, a simple computation shows that $\frac{2\lambda}{\mu(\gamma)} \leq \frac{2\lambda}{\mu(\frac{2}{21})} < 91$. Since $\frac{2\lambda(2+\mu(\gamma))}{3}$ is a decreasing function of γ, we also have $\frac{2\lambda(2+\mu(\gamma))}{3} \leq \frac{2\lambda(2+\mu(0))}{3} < \frac{5}{2}$. This proves (68.9) for $n = 1$. Let us now consider the case $n \geq 2$. Since we have $\|2u_h^n - u_h^{n-1}\|_L^2 \leq 6\|u_h^n\|_L^2 + 3\|u_h^{n-1}\|_L^2$, we infer that $\|u_h^n\|_L^2 + \|2u_h^n - u_h^{n-1}\|_L^2 \leq 7\|u_h^n\|_L^2 + 3\|u_h^{n-1}\|_L^2$. This in turn implies that

$$\frac{1}{7}\frac{\alpha}{\iota_{L,V}^2}\big(\|u_h^n\|_L^2 + \|2u_h^n - u_h^{n-1}\|_L^2\big) \leq \alpha\|u_h^n\|_V^2 + \frac{3}{7}\alpha\|u_h^{n-1}\|_V^2.$$

Recalling that $\frac{2}{\rho} = \frac{\alpha}{\iota_{L,V}^2}$ and $\gamma = \frac{2}{7}\frac{\tau}{\rho}$, the stability estimate (68.7) gives

$$(1+\gamma)\big(\|u_h^n\|_L^2 + \|2u_h^n - u_h^{n-1}\|_L^2\big) + \alpha\tau\|u_h^n\|_V^2 \\ \leq \|u_h^{n-1}\|_L^2 + \|2u_h^{n-1} - u_h^{n-2}\|_L^2 + \frac{3}{7}\alpha\tau\|u_h^{n-1}\|_V^2 + \frac{2}{\alpha}\tau\|f^n\|_{V'}^2.$$

The assumption $\gamma \leq \frac{2}{21}$ implies that $\frac{3}{7}\alpha\tau \leq \frac{\alpha\tau}{1+\gamma}$. This proves (68.9) for $n \geq 2$. Having established that (68.9) holds true for all $n \in \mathcal{N}_\tau$, we obtain the expected bound by invoking the incremental Gronwall lemma from Exercise 67.1 and by observing that $(1+\gamma)^{-1} \leq e^{-\frac{7\gamma}{8}} = e^{-\frac{\tau}{4\rho}}$ for $\gamma \in (0, \frac{2}{21})$ (see also the proof of Lemma 67.5). □

Remark 68.3 (Literature). The BDF2 scheme belongs to the class of multistep schemes. The analysis of these schemes was started among others by Zlámal [294], Crouzeix and Raviart [95], Crouzeix [94]. We also refer the reader to Thomée [273, Chap. 10] and the references therein. The proof of Lemma 68.1 somewhat differs from the argument in [273, Thm. 1.7] (see also Exercise 68.4) which combines the stability argument with the error estimate. Here, (68.7) has a telescoping form and delivers a bound on the discrete second-order time derivative $\|u_h^n - 2u_h^{n-1} + u_h^{n-2}\|_L$. □

68.2.3 Error analysis

Error bounds with V'-valued time derivatives of the solution to (65.10) can be derived by proceeding as in Theorem 67.6, but to avoid the proliferation of estimates, we are just going to invoke the embedding $L \hookrightarrow V'$ to estimate the error bounds with L-valued time derivatives. Let us start by estimating the error in the $\ell^2(J;V)$-norm. Recall that α and M denote the coercivity and boundedness constants of the bilinear form a and $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$ is a time scale. As for the implicit Euler scheme, we consider a generic operator $\mathcal{I}_h : V \to V_h$ having optimal approximation properties.

Theorem 68.4 ($\ell^2(J;V)$-estimate). *Let u solve (65.10) and assume that $u \in H^1(J;V) \cap H^3(J;L)$. Let $\eta(t) := u(t) - \mathcal{I}_h(u(t))$ for all $t \in J$. Set $u_\tau := (u(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$ and $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau} \in (V)^N$. Let $u_{h\tau} \in (V_h)^N$ solve (68.1). There is c s.t. for all $h \in \mathcal{H}$, τ, α, and M,*

$$\|u_\tau - u_{h\tau}\|_{\ell^2(J;V)} \le c\left(\tau^2\Big(\frac{1}{\sqrt{\alpha}}\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)} + \frac{\rho}{\iota_{L,V}}\|\partial_{ttt}u\|_{L^2(J;L)}\Big)\right.$$
$$\left. + \Big(1+\frac{M}{\alpha}\Big)\|\eta_\tau\|_{\ell^2(J;V)} + \frac{\rho}{\iota_{L,V}}\|\partial_t\eta\|_{L^2(J;L)} + \frac{1}{\sqrt{\alpha}}\|\eta(0)\|_L\right). \tag{68.10}$$

Proof. (1) Using the usual notation $e_h^n := u_h^n - \mathcal{I}_h(u(t_n))$, we have for all $w_h \in V_h$,

$$(e_h^1 - e_h^0, w_h)_L + \tau a^1(e_h^1, w_h) = \tau\langle g^1, w_h\rangle_{V',V},$$

with $\langle g^1, w_h\rangle_{V',V} := a^1(\eta(t_1), w_h) + \langle \xi^1 - \psi^1, w_h\rangle_{V',V}$, $\xi^1 := \frac{1}{\tau}\int_{J_1}\partial_t\eta(t)\mathrm{d}t$, and $\psi^1 := \frac{1}{\tau}\int_{J_1}(t_0 - t)\partial_{tt}u(t)\mathrm{d}t$ (see the proof of Theorem 67.6). The key idea to achieve optimality in the error estimate is to split g^1. We set $g^1 := g^1_{V'} + g^1_L$, with $g^1_{V'} \in V'$ and $g^1_L \in L$, where $\langle g^1_{V'}, w_h\rangle_{V',V} := a^1(\eta(t_1), w_h) + \langle \xi^1, w_h\rangle_{V',V}$ for all $w_h \in V_h$, and $g^1_L := -\psi^1$. This implies that

$$(e_h^1 - e_h^0, w_h)_L + \tau a^1(e_h^1, w_h) = \tau\langle g^1_{V'}, w_h\rangle_{V',V} + \tau(g^1_L, w_h)_L. \tag{68.11}$$

The triangle inequality, the boundedness of a^1, and the Cauchy–Schwarz inequality imply that $\|g^1_{V'}\|_{V'} \le M\|\eta(t_1)\|_V + \tau^{-\frac{1}{2}}\|\partial_t\eta\|_{L^2(J_1;V')}$, and since $\|\phi\|_{V'} \le \iota_{L,V}\|\phi\|_L$ for all $\phi \in L$, we infer that

$$\|g^1_{V'}\|_{V'} \le M\|\eta(t_1)\|_V + \tau^{-\frac{1}{2}}\iota_{L,V}\|\partial_t\eta\|_{L^2(J_1;L)}.$$

Moreover, we have $\|g^1_L\|_L \le \tau\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)}$. Notice that $u \in H^3(J;L)$ implies that $u \in C^2(\overline{J}_1;L)$.
(2) We have for all $n \in \mathcal{N}_\tau$, $n \ge 2$,

$$(\tfrac{3}{2}e_h^n - 2e_h^{n-1} + \tfrac{1}{2}e_h^{n-2}, w_h)_L + \tau a^n(e_h^n, w_h) = \tau\langle g^n, w_h\rangle_{V',V}, \tag{68.12}$$

with $\langle g^n, w_h\rangle_{V',V} := a^n(\eta(t_n), w_h) + \langle \xi^n - \psi^n, w_h\rangle_{V',V}$, $\xi^n := \tau^{-1}(\frac{3}{2}\eta(t_n) - 2\eta(t_{n-1}) + \frac{1}{2}\eta(t_{n-2}))$, and $\psi^n := \tau^{-1}(\frac{3}{2}u(t_n) - 2u(t_{n-1}) + \frac{1}{2}u(t_{n-2}) - \tau\partial_t u(t_n))$. A direct calculation shows that

$$\xi^n = \frac{3}{2\tau}\int_{J_n}\partial_t\eta(t)\mathrm{d}t - \frac{1}{2\tau}\int_{J_{n-1}}\partial_t\eta(t)\mathrm{d}t,$$
$$\psi^n = \frac{1}{\tau}\int_{J_n}(t - t_{n-1})^2\partial_{ttt}u(t)\mathrm{d}t - \frac{1}{4\tau}\int_{J_{n-1}\cup J_n}(t - t_{n-2})^2\partial_{ttt}u(t)\mathrm{d}t.$$

The triangle inequality and the Cauchy–Schwarz inequality imply that

$$\|\xi^n\|_{V'} \le \tau^{-1}\Big(\frac{3}{2}\tau^{\frac{1}{2}}\|\partial_t\eta\|_{L^2(J_n;V')} + \frac{1}{2}\tau^{\frac{1}{2}}\|\partial_t\eta\|_{L^2(J_{n-1};V')}\Big)$$
$$\le \tau^{-\frac{1}{2}}\big(\tfrac{5}{2}\big)^{\frac{1}{2}}\|\partial_t\eta\|_{L^2(J_{n-1}\cup J_n;V')}.$$

Similarly, we have

$$\begin{aligned}\|\psi^n\|_{V'} &\le \tau^{-1}\Big(\big(\tfrac{1}{5}\big)^{\frac12}\tau^{\frac52}\|\partial_{ttt}u\|_{L^2(J_n;V')} + \big(\tfrac{2}{5}\big)^{\frac12}\tau^{\frac52}\|\partial_{ttt}u\|_{L^2(J_{n-1}\cup J_n;V')}\Big)\\ &\le \tau^{\frac32}\big(\tfrac{6}{5}\big)^{\frac12}\|\partial_{ttt}u\|_{L^2(J_{n-1}\cup J_n;V')},\end{aligned}$$

since $(\frac{1}{5})^{\frac12} + (\frac{2}{5})^{\frac12} \le \big(\frac{6}{5}\big)^{\frac12}$. Invoking the triangle inequality, the boundedness of a^n, and the boundedness of the embedding $V \hookrightarrow L$ implies that

$$\begin{aligned}\|g^n\|_{V'} \le M\|\eta(t_n)\|_V &+ \big(\tfrac{5}{2}\big)^{\frac12}\tau^{-\frac12}\iota_{L,V}\|\partial_t\eta\|_{L^2(J_{n-1}\cup J_n;L)}\\ &+ \big(\tfrac{6}{5}\big)^{\frac12}\tau^{\frac32}\iota_{L,V}\|\partial_{ttt}u\|_{L^2(J_{n-1}\cup J_n;L)}.\end{aligned}$$

(3) We now adapt the proof of Theorem 67.6. The stability estimate (68.4) established in Lemma 68.1 implies that

$$\alpha\|e_{h\tau}\|^2_{\ell^2(J;V)} \le \frac{2}{\alpha}\|\tilde g_\tau\|^2_{\ell^2(J;V')} + 29\tau^2\|g_L^1\|_L^2 + \frac52\|e_h^0\|_L^2, \tag{68.13}$$

with $\tilde g^1 := g^1_{V'}$ and $\tilde g^n := g^n$ for all $n \in \mathcal{N}_\tau$, $n \ge 2$. Using the above bounds on $\|g^1_{V'}\|_{V'}$ and $\|g^n\|_{V'}$ for all $n \in \mathcal{N}_\tau$, $n \ge 2$, we infer that

$$\begin{aligned}\|\tilde g_\tau\|^2_{\ell^2(J;V')} &= \sum_{n\in\mathcal{N}_\tau}\tau\|\tilde g^n\|^2_{V'} = \tau\|g^1_{V'}\|^2_{V'} + \sum_{n\in\mathcal{N}_\tau,n\ge2}\tau\|g^n\|^2_{V'}\\ &\le \tau\big(M\|\eta(t_1)\|_V + \tau^{-\frac12}\iota_{L,V}\|\partial_t\eta\|_{L^2(J_1;L)}\big)^2 + \sum_{n\in\mathcal{N}_\tau,n\ge2}\tau\Big(M\|\eta(t_n)\|_V\\ &\quad + \big(\tfrac{5}{2}\big)^{\frac12}\tau^{-\frac12}\iota_{L,V}\|\partial_t\eta\|_{L^2(J_{n-1}\cup J_n;L)} + \big(\tfrac{6}{5}\big)^{\frac12}\tau^{\frac32}\iota_{L,V}\|\partial_{ttt}u\|_{L^2(J_{n-1}\cup J_n;L)}\Big)^2\\ &\le c\sum_{n\in\mathcal{N}_\tau}\Big(\tau M^2\|\eta(t_n)\|_V^2 + \iota^2_{L,V}\|\partial_t\eta\|^2_{L^2(J_n;L)} + \tau^4\iota^2_{L,V}\|\partial_{ttt}u\|^2_{L^2(J_n;L)}\Big)\\ &= c\big(M^2\|\eta_\tau\|^2_{\ell^2(J;V)} + \iota^2_{L,V}\|\partial_t\eta\|^2_{L^2(J;L)} + \tau^4\iota^2_{L,V}\|\partial_{ttt}u\|^2_{L^2(J;L)}\big).\end{aligned}$$

Using this bound together with the above estimate on $\|g_L^1\|_L$ on the right-hand side of (68.13) and dividing by $\alpha > 0$ leads to

$$\begin{aligned}\|e_{h\tau}\|^2_{\ell^2(J;V)} \le c\Big(&\frac{M^2}{\alpha^2}\|\eta_\tau\|^2_{\ell^2(J;V)} + \frac{\iota^2_{L,V}}{\alpha^2}\|\partial_t\eta\|^2_{L^2(J;L)} + \frac{\iota^2_{L,V}}{\alpha^2}\tau^4\|\partial_{ttt}u\|^2_{L^2(J;L)}\\ &+ \frac{1}{\alpha}\tau^4\|\partial_{tt}u\|^2_{C^0(\overline{J}_1;L)} + \frac{1}{\alpha}\|e_h^0\|_L^2\Big).\end{aligned}$$

Taking the square root and rearranging the terms, we infer that

$$\begin{aligned}\|e_{h\tau}\|_{\ell^2(J;V)} \le c\Big(&\frac{1}{\sqrt{\alpha}}\|\eta(0)\|_L + \frac{M}{\alpha}\|\eta_\tau\|_{\ell^2(J;V)} + \frac{\rho}{\iota_{L,V}}\|\partial_t\eta\|_{L^2(J;L)}\\ &+ \frac{1}{\sqrt{\alpha}}\tau^2\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)} + \frac{\rho}{\iota_{L,V}}\tau^2\|\partial_{ttt}u\|_{L^2(J;L)}\Big),\end{aligned} \tag{68.14}$$

where we used that $\rho := 2\frac{\iota^2_{L,V}}{\alpha}$ and $\|e_h^0\|_L \le \|\eta(0)\|_L$ (which results from $u_h^0 := \mathcal{P}_{V_h}(u_0)$). We obtain (68.10) by invoking the triangle inequality on $u_h^n - u(t_n) = e_h^n - \eta(t_n)$. □

Remark 68.5 (Initial split $g^1 = g^1_{V'} + g^1_L$). Observe that the error estimate (68.10) scales optimally like $\mathcal{O}(\tau^2)$ although the first time step is only first-order accurate in time. The special treatment we gave to the consistency tem g^1 is the key to obtain optimality. More precisely, since the consistency term $\|g^1_L\|^2_L$ in (68.13) is multiplied by τ^2, the corresponding error $\tau^4\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)}$ scales optimally with respect to τ. □

Remark 68.6 (Supercloseness). Assuming that the bilinear form a is time-independent for simplicity, another interesting choice for the error decomposition is to set $\eta(t) := u(t) - \Pi^{\mathrm{E}}_h(u(t))$, where $\Pi^{\mathrm{E}}_h : V \to V_h$ the elliptic projection defined in (66.14) (that is, $a(\Pi^{\mathrm{E}}_h(v), w_h) := a(v, w_h)$ for all $v \in V$ and all $w_h \in V_h$). This leads to a supercloseness estimate on the discrete error $e_{h\tau}$, where the term $\frac{M}{\alpha}\|\eta_\tau\|_{\ell^2(J;V)}$ disappears from the upper bound in (68.14) (see Remark 67.8 for a similar result for the implicit Euler scheme). □

We now derive an error estimate in the $\ell^\infty(\overline{J};L)$-norm. The improvement with respect to Theorem 68.4 is twofold. On the one hand we capture the exponential decay of the influence of the initial errors. On the other hand the use of the elliptic projection allows us to avoid estimating the error using the $\ell^2(J;V)$-norm.

Theorem 68.7 (Improved $\ell^\infty(\overline{J};L)$-estimate). *In addition to the hypotheses of Theorem* 68.4, *assume that $\tau \le \frac{1}{3}\rho$ and that the bilinear form a is time-independent. Let $\eta(t) := u(t) - \Pi^{\mathrm{E}}_h(u(t))$, where $\Pi^{\mathrm{E}}_h : V \to V_h$ is the elliptic projection defined in* (66.14). *There are c_1, c_2 s.t. for all $h \in \mathcal{H}$, τ, α, and M, we have for all $n \in \mathcal{N}_\tau$,*

$$\begin{aligned}\|u^n_h - u(t_n)\|_L \le \|\eta(t_n)\|_L + c_1\left(e^{-\frac{t_n}{8\rho}}\|\eta(0)\|_L + \sqrt{\rho}\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_t\eta\|_{L^2((0,t_n);L)}\right)\\ + c_2\,\tau^2\left(e^{-\frac{t_n}{8\rho}}\|\partial_{tt}u\|_{C^0(\overline{J}_1;L)} + \sqrt{\rho}\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_{ttt}u\|_{L^2((0,t_n);L)}\right).\end{aligned} \tag{68.15}$$

Proof. We set $v_h(t) := \Pi^{\mathrm{E}}_h(u(t))$ in (68.11) and (68.12), i.e., $e^n_h := u_h - \Pi^{\mathrm{E}}_h(u(t_n))$ and $\eta(t) := u(t) - \Pi^{\mathrm{E}}_h(u(t))$. This implies that we now have $g^1_{V'} := \xi^1$, $g^1_L := -\psi^1$ (as before), and $g^n := \xi^n - \psi^n$ for all $n \in \mathcal{N}_\tau$, $n \ge 2$. The stability estimate (68.8) established in Lemma 68.2 becomes

$$\|e^n_h\|^2_L \le e^{-\frac{t_n}{4\rho}}\left(91\tau^2\|g^1_L\|^2_L + \frac{5}{2}\|e^0_h\|^2_L\right) + \frac{2}{\alpha}\tau\sum_{k\in\{1:n\}} e^{-\frac{t_n-t_{k-1}}{4\rho}}\|\tilde{g}^k\|^2_{V'},$$

with $\tilde{g}^1 := g^1_{V'}$ and $\tilde{g}^n := g^n$ for all $n \in \mathcal{N}_\tau$, $n \ge 2$. Using the bounds on on $\|g^1_{V'}\|_{V'}$, $\|g^1_L\|_L$, and $\|g^n\|_{V'}$ derived in the previous proof, we infer that there is c s.t. for all $h \in \mathcal{H}$, τ, α, and M,

$$\begin{aligned}\|e^n_h\|^2_L \le e^{-\frac{t_n}{4\rho}}\left(91\tau^4\|\partial_{tt}u\|^2_{C^0(\overline{J}_1;L)} + \frac{5}{2}\|e^0_h\|^2_L\right)\\ + c\,\frac{1}{\alpha}\sum_{k\in\{1:n\}} e^{-\frac{t_n-t_{k-1}}{4\rho}}\iota^2_{L,V}(\|\partial_t\eta\|^2_{L^2(J_k;L)} + \tau^4\|\partial_{ttt}u\|^2_{L^2(J_k;L)}),\end{aligned}$$

where we used that $e^{-\frac{t_n-t_{k-1}}{4\rho}} \le e^{\frac{1}{12}}e^{-\frac{t_n-t_{k-2}}{4\rho}}$ since $\tau \le \frac{1}{3}\rho$. Moreover, since $e^{-\frac{t_n-t_{k-1}}{4\rho}} \le e^{-\frac{t_n-s}{4\rho}}$ for all $s \in J_k$, and recalling that $\rho := 2\frac{\iota^2_{L,V}}{\alpha}$ and $\|e^0_h\|_L \le \|\eta(0)\|_L$, we obtain

$$\begin{aligned}\|e^n_h\|^2_L \le e^{-\frac{t_n}{4\rho}}\left(91\tau^4\|\partial_{tt}u\|^2_{C^0(\overline{J}_1;L)} + \frac{5}{2}\|\eta(0)\|^2_L\right)\\ + c\,\rho\left(\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_t\eta\|^2_{L^2((0,t_n);L)} + \tau^4\|e^{-\frac{t_n-\cdot}{8\rho}}\partial_{ttt}u\|^2_{L^2((0,t_n);L)}\right).\end{aligned}$$

Taking the square root, invoking the triangle inequality on $u^n_h - u(t_n) = e^n_h - \eta(t_n)$, and rearranging the terms proves the assertion. □

Example 68.8 (Heat equation). Let us consider the approximation of the heat equation with H^1-conforming finite elements and BDF2. Let $r \in [1,k]$, where $k \geq 1$ is the degree of the finite elements used to build the discrete space V_h. Assume that $u \in C^0(\overline{J};H^{r+1}(D)) \cap H^1(J;H^r(D)) \cap H^3(J;L^2(D))$. Then Theorem 68.4 implies that the error in the $\ell^2(J;H_0^1(D))$-norm decays as $\mathcal{O}(h^r+\tau^2)$ and if $u \in H^1(J;H^{r+1}(D)) \cap H^3(J;L^2(D))$, Theorem 68.7 implies that the error in the $\ell^\infty(\overline{J};L^2(D))$-norm decays as $\mathcal{O}(h^{r+s}+\tau^2)$, where $s \in (0,1]$ is the elliptic regularity pickup index ($s=1$ if there is full elliptic regularity pickup). □

68.3 Crank–Nicolson scheme

We review in this section a method introduced in [93, Eq. (5)] which is now known in the literature as the *Crank–Nicolson scheme*. This scheme is, as the implicit Euler scheme, a one-step method.

68.3.1 Principle and algebraic realization

The Crank–Nicolson scheme is based on the midpoint rule $\partial_t u_h(t_{n-\frac12}) = \frac{1}{\tau}(u_h(t_n)-u_h(t_{n-1})) + \mathcal{O}(\tau^2)$, where $t_{n-\frac12} := t_{n-1}+\frac{\tau}{2}$. After setting $u_h^0 := \mathcal{P}_{V_h}(u_0)$, as for the Euler schemes, we construct the sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ such that

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a^{n-\frac12}(\tfrac12(u_h^n+u_h^{n-1}), w_h) = \tau\langle f^{n-\frac12}, w_h\rangle_{V',V}, \tag{68.16}$$

for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, with $a^{n-\frac12}(\cdot,\cdot) := a(t_{n-\frac12};\cdot,\cdot)$ and $f^{n-\frac12} := f(t_{n-\frac12}) \in V'$.

Let U^n, U^{n-1} be the coordinate vectors of u_h^n and u_h^{n-1} in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, respectively. Consider the stiffness matrix $\mathcal{A}^{n-\frac12} \in \mathbb{R}^{I\times I}$ s.t. $\mathcal{A}_{ij}^{n-\frac12} := a(t_{n-\frac12};\varphi_j,\varphi_i)$ and recall the mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$ s.t. $\mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L$ for all $i,j \in \{1:I\}$. Then the algebraic realization of the Crank–Nicolson scheme is

$$\mathcal{M}\mathsf{U}^n + \frac12\tau\mathcal{A}^{n-\frac12}\mathsf{U}^n = \mathcal{M}\mathsf{U}^{n-1} - \frac12\tau\mathcal{A}^{n-\frac12}\mathsf{U}^{n-1} + \tau\mathsf{F}^{n-\frac12}, \tag{68.17}$$

with $\mathsf{F}^{n-\frac12} := (\langle f^{n-\frac12},\varphi_i\rangle_{V',V})_{i\in\{1:I\}}$. This expression shows that the computational cost of one step of the Crank–Nicolson scheme is comparable to that of the implicit Euler scheme.

68.3.2 Stability

We start by establishing some stability estimates using the coercivity argument. Recall that α denotes the coercivity constant of the bilinear form a. As above, we consider the time-discrete norm $\|\phi_\tau\|_{\ell^2(J;B)}^2 := \sum_{n\in\mathcal{N}_\tau}\tau\|\phi^n\|_B^2$ with $\phi_\tau := (\phi^n)_{n\in\mathcal{N}_\tau} \in B^N$, $B := V$ or $B := V'$ (see (67.1)). Moreover, for every sequence $\tilde v_{h\tau} := (v_h^0, v_{h\tau}) \in (V_h)^{N+1}$, we set $\bar v_h^n := \frac12(v_h^n + v_h^{n-1})$ for all $n \in \mathcal{N}_\tau$. Denoting $u_{h\tau} \in (V_h)^N$ the solution to (68.16), we set $\tilde u_{h\tau} := (u_h^0, u_{h\tau})$ and define $\bar u_h^n$ accordingly.

Lemma 68.9 ($\ell^2(J;V)$-stability). *Let $u_{h\tau} \in (V_h)^N$ solve (68.16) with the sequence of source terms $f_\tau := (f^{n-\frac12})_{n\in\mathcal{N}_\tau} \in (V')^N$. The following holds true:*

$$\alpha\|\bar u_{h\tau}\|_{\ell^2(J;V)}^2 + \|u_h^N\|_L^2 \leq \frac{1}{\alpha}\|f_\tau\|_{\ell^2(J;V')}^2 + \|u_h^0\|_L^2. \tag{68.18}$$

Proof. Using $w_h := \frac{1}{2}(u_n^n + u_h^{n-1}) =: \bar{u}_h^n$ as the test function in (68.16), observing that $(u_h^n - u_h^{n-1}, \bar{u}_h^n)_L = \frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2$, and employing Young's inequality to bound $\langle f^{n-\frac{1}{2}}, \bar{u}_h^n\rangle_{V',V}$, we obtain

$$\frac{1}{2}\|u_h^n\|_L^2 + \frac{1}{2}\alpha\tau\|\bar{u}_n^n\|_V^2 \le \frac{1}{2}\|u_h^{n-1}\|_L^2 + \frac{1}{2\alpha}\tau\|f^{n-\frac{1}{2}}\|_{V'}^2.$$

We conclude by summing the above inequality over $n \in \mathcal{N}_\tau$. □

Remark 68.10 (Comparison). The stability estimate (68.18) only controls the $\ell^2(J;V)$-norm of $\bar{u}_{h\tau}$, whereas the implicit Euler scheme and the BDF2 scheme both control the $\ell^2(J;V)$-norm of $u_{h\tau}$ (see Lemma 67.3 and Lemma 68.1, respectively). Notice that the left-hand side of (68.18) still defines a norm on $u_{h\tau}$ since $\alpha\|\bar{u}_{h\tau}\|_{\ell^2(J;V)}^2 + \|u_h^N\|_L^2 = 0$ implies that $u_h^N = 0$, $\bar{u}_h^N = 0$, $u_h^{N-1} = 2u_h^N - \bar{u}_h^N = 0$, $\bar{u}_h^{N-1} = 0$, and so on until $u_h^0 = 0$. Moreover, it is also possible to establish an inf-sup condition for the bilinear form associated with the Crank–Nicolson scheme in the spirit of what was done in Exercise 67.2 for the implicit Euler scheme. We do not detail this result here for brevity since it is a particular case of the inf-sup condition established in Lemma 71.20 for the more general class of continuous Petrov–Galerkin schemes of arbitrary order (the Crank–Nicolson scheme is the lowest-order scheme in this class). □

Remark 68.11 (θ-schemes). The implicit Euler, the Crank–Nicolson, and the explicit Euler schemes are part of a family of methods parameterized by $\theta \in [0,1]$ which approximate the bilinear form a over the time interval J_n as $a(t_{n-1}+\theta\tau, (1-\theta)u_h^{n-1} + \theta u_h^n, w_h)$ and the time derivative by $\frac{1}{\tau}(u_h^{n+1} - u_h^n)$. This leads to

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a^{n-1+\theta}((1-\theta)u_h^{n-1} + \theta u_h^n, w_h) = \langle f^{n-1+\theta}, w_h\rangle_{V',V},$$

for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, with $a^{n-1+\theta}(\cdot,\cdot) := a(t_{n-1}+\theta\tau;\cdot,\cdot)$ and $f^{n-1+\theta} := f(t_{n-1}+\theta\tau) \in V'$. The θ-scheme can be shown to be unconditionally stable when $\theta \in (\frac{1}{2}, 1]$ and conditionally stable when $\theta \in [0, \frac{1}{2})$. The method corresponding to $\theta = \frac{1}{2}$, which is the Crank–Nicolson scheme, is said to be *marginally stable*. Notice that the Crank–Nicolson scheme is the only one in this family that is second-order accurate in time. □

We now establish a sharper stability estimate in the $\ell^\infty(\overline{J};L)$-norm that captures the exponentially decaying influence of the data on the solution. However, contrary to the implicit Euler and BDF2 schemes, this sharper estimate hinges on some assumptions on the data and on the time step. In particular, the bound on the time step involves the following mesh-dependent parameter already introduced in §67.2.2 for the analysis of the explicit Euler method:

$$c_{\mathrm{INV}}(h) := \iota_{L,V} \max_{v_h \in V_h} \frac{\|v_h\|_V}{\|v_h\|_L}. \tag{68.19}$$

This quantity is nondimensional and it is finite since V_h is finite-dimensional. For the heat equation, we have $V := H_0^1(D)$, $L := L^2(D)$, with $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$ and $\|v\|_L := \|v\|_{L^2(D)}$, and in this context we have $\iota_{L,V} := C_{\mathrm{PS}}^{-1}\ell_D$, where C_{PS} is the Poincaré–Steklov constant in $H_0^1(D)$ and ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. Assuming that V_h is a finite element space based on a quasi-uniform mesh sequence, the inverse inequality in Lemma 12.1 shows that $c_{\mathrm{INV}}(h) \le c\ell_D h^{-1}$ for all $h \in \mathcal{H}$.

Lemma 68.12 ($\ell^\infty(\overline{J};L)$-stability, exponential decay). *Assume that $f \in C^0(\overline{J};L)$ and that the bilinear form a is time-independent. Let $u_{h\tau} \in (V_h)^N$ solve* (68.16) *with the sequence of source terms $f_\tau := (f^{n-\frac{1}{2}})_{n\in\mathcal{N}_\tau} \in (L)^N$. Assume that*

$$\tau \le \frac{\rho}{2}\min\left(1, \xi_\kappa^{-1} c_{\mathrm{INV}}(h)^{-1}\right), \tag{68.20}$$

with $\xi_\kappa := \frac{M}{\alpha}$. *The following holds true for all* $n \in \mathcal{N}_\tau$*:*

$$\|u_h^n\|_L^2 \le e^{-\frac{t_n}{\rho}}(\|u_h^0\|_L^2 + \frac{\rho\alpha}{8}\|u_h^0\|_V^2) + \frac{7\rho}{2}\sum_{k\in\{1:n\}}\tau e^{-\frac{t_n-t_{k-1}}{\rho}}\|f^{k-\frac12}\|_L^2. \tag{68.21}$$

Proof. For all $n \in \mathcal{N}_\tau$, let us define the linear operator $A_h : V_h \to V_h$ by setting $(A_h(v_h), w_h)_L := a(v_h, w_h)$ for all $v_h, w_h \in V_h$. Let us set $f_h^{n-\frac12} := \mathcal{P}_{V_h}(f^{n-\frac12})$ and notice that by assumption we have $\|f_h^{n-\frac12}\|_L \le \|f^{n-\frac12}\|_L$. The Crank–Nicolson scheme (68.16) can be rewritten as follows:

$$u_h^n + \frac12\tau A_h(u_h^n) = u_h^{n-1} - \frac12\tau A_h(u_h^{n-1}) + f_h^{n-\frac12}.$$

Squaring this equality and developing the squares, we obtain

$$\begin{aligned}
&\|u_h^n\|_L^2 + \tau a(u_h^n, u_h^n) + \frac14\tau^2\|A_h(u_h^n)\|_L^2\\
&\quad= \|u_h^{n-1}\|_L^2 - \tau a(u_h^{n-1}, u_h^{n-1}) + \frac14\tau^2\|A_h(u_h^{n-1})\|_L^2\\
&\qquad+ 2\tau(u_h^{n-1}, f_h^{n-\frac12})_L - \tau^2(A_h(u_h^{n-1}), f_h^{n-\frac12})_L + \tau^2\|f_h^{n-\frac12}\|_L^2.
\end{aligned}$$

Using the coercivity of a on the left- and right-hand sides, we infer that

$$\begin{aligned}
&\|u_h^n\|_L^2 + \alpha\tau\|u_h^n\|_V^2 + \frac{\tau^2}{4}\|A_h(u_h^n)\|_L^2\\
&\quad\le \|u_h^{n-1}\|_L^2 - \alpha\tau\|u_h^{n-1}\|_V^2 + \frac{\tau^2}{4}\|A_h(u_h^{n-1})\|_L^2\\
&\qquad+ 2\tau(u_h^{n-1}, f_h^{n-\frac12})_L - \tau^2(A_h(u_h^{n-1}), f_h^{n-\frac12})_L + \tau^2\|f_h^{n-\frac12}\|_L^2.
\end{aligned}$$

We now estimate the third, fourth, and fifth terms on the right-hand side. For the third term, we use the bound (see Exercise 68.2)

$$\|A_h(v_h)\|_L \le \iota_{L,V}^{-1}c_{\mathrm{INV}}(h)M\|v_h\|_V, \qquad \forall v_h \in V_h, \tag{68.22}$$

applied with $v_h := u_h^{n-1}$ and proceed as follows:

$$\begin{aligned}
\frac{\tau^2}{4}\|A_h(u_h^{n-1})\|_L^2 &= \frac{\frac{\tau^2}{4}}{1+\frac{2\tau}{\rho}}\|A_h(u_h^{n-1})\|_L^2 + \frac{\frac{\tau^3}{2\rho}}{1+\frac{2\tau}{\rho}}\|A_h(u_h^{n-1})\|_L^2\\
&\le \frac{\frac{\tau^2}{4}}{1+\frac{2\tau}{\rho}}\|A_h(u_h^{n-1})\|_L^2 + \frac{\tau^3}{2\rho}\|A_h(u_h^{n-1})\|_L^2\\
&\le \frac{\frac{\tau^2}{4}}{1+\frac{2\tau}{\rho}}\|A_h(u_h^{n-1})\|_L^2 + \frac14\alpha\tau\|u_h^{n-1}\|_V^2,
\end{aligned}$$

where we used that $\frac{\tau^3}{2\rho}\iota_{L,V}^{-2}c_{\mathrm{INV}}(h)^2M^2 = \alpha\tau^3\rho^{-2}c_{\mathrm{INV}}(h)^2\xi_\kappa^2 \le \frac14\alpha\tau$ owing to the condition (68.20) on the time step. For the fourth term, we use the Cauchy–Schwarz inequality, the embedding inequality $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$, and Young's inequality to infer that

$$\begin{aligned}
2\tau\big|(u_h^{n-1}, f_h^{n-\frac12})_L\big| &\le 2\tau\|u_h^{n-1}\|_L\|f_h^{n-\frac12}\|_L \le 2\tau\iota_{L,V}\|u_h^{n-1}\|_V\|f_h^{n-\frac12}\|_L\\
&\le \tau\frac{\iota_{L,V}^2}{\rho}\|u_h^{n-1}\|_V^2 + \tau\rho\|f_h^{n-\frac12}\|_L^2\\
&\le \frac12\alpha\tau\|u_h^{n-1}\|_V^2 + \tau\rho\|f_h^{n-\frac12}\|_L^2.
\end{aligned}$$

For the fifth term, we use the Cauchy–Schwarz inequality, the above bound (68.22) applied with $v_h := u_h^{n-1}$, and Young's inequality to obtain

$$\begin{aligned}\left|\tau^2(A_h(u_h^{n-1}), f_h^{n-\frac12})_L\right| &\le \frac{\tau^3}{2\rho}\|A_h(u_h^{n-1})\|_L^2 + 2\tau\rho\|f_h^{n-\frac12}\|_L^2\\ &\le \frac14\alpha\tau\|u_h^{n-1}\|_V^2 + 2\tau\rho\|f_h^{n-\frac12}\|_L^2,\end{aligned}$$

where the last bound uses the same arguments as above. Putting everything together, we infer that

$$\begin{aligned}&\|u_h^n\|_L^2 + \alpha\tau\|u_h^n\|_V^2 + \frac{\tau^2}{4}\|A_h(u_h^n)\|_L^2\\ &\qquad\le \|u_h^{n-1}\|_L^2 + \frac{\frac{\tau^2}{4}}{1+\frac{2\tau}{\rho}}\|A_h(u_h^{n-1})\|_L^2 + (3\tau\rho+\tau^2)\|f_h^{n-\frac12}\|_L^2.\end{aligned}$$

Letting $\gamma := \frac{2\tau}{\rho}$ and invoking one more time the embedding $V \hookrightarrow L$, we have

$$(1+\gamma)\|u_h^n\|_L^2 \le \|u_h^n\|_L^2 + \alpha\tau\|u_h^n\|_V^2,$$

so that the above estimate can be rewritten as $(1+\gamma)a_n \le a_{n-1} + b_n$ with $a_n := \|u_h^n\|_L^2 + \frac{\tau^2}{4(1+\gamma)}\|A_h(u_h^n)\|_L^2$ and $b_n := \frac72\tau\rho\|f_h^{n-\frac12}\|_L^2$ (note that $3\tau\rho+\tau^2 \le \frac72\tau\rho$ since $\tau \le \frac12\rho$ by assumption). Invoking the incremental Gronwall lemma from Exercise 67.1 and since $a_0 \le \|u_h^0\|_L^2 + \frac{\tau^2}{4}\|A_h(u_h^0)\|_L^2 \le \|u_h^0\|_L^2 + \frac{\alpha\rho}{8}\|u_h^0\|_V^2$ (where this last bound is again a consequence of (68.22) and the restriction (68.20) on the time step) leads to

$$\|u_h^n\|_L^2 \le \frac{1}{(1+\frac{2\tau}{\rho})^n}(\|u_h^0\|_L^2 + \frac{\rho\alpha}{8}\|u_h^0\|_V^2) + \frac{7\rho}{2}\sum_{k\in\{1:n\}}\tau\frac{\|f^{k-\frac12}\|_L^2}{(1+\frac{2\tau}{\rho})^{n-k+1}}.$$

The assertion follows by recalling that $\frac{2\tau}{\rho} \le 1$ yields $(1+\frac{2\tau}{\rho})^{-1} \le e^{-\frac{\tau}{\rho}}$. □

Remark 68.13 (Assumptions). For the heat equation, the condition $\tau \le \frac{\rho}{2}\xi_\kappa^{-1}c_{\mathrm{INV}}(h)^{-1}$ leads to an upper bound on the time step proportional to h. This is significantly less restrictive than the parabolic CFL required for the explicit Euler scheme which imposes an upper bound on the time step proportional to h^2. Moreover, the assumption that the bilinear form a is time-independent can be lifted by slightly modifying the Crank–Nicolson scheme (see Exercise 68.5). □

Remark 68.14 (Literature). The analysis of the Crank–Nicolson scheme was started among others by Douglas and Dupont [109], Baker et al. [21], Douglas et al. [110]. We also refer the reader to Thomée [273, Thm. 1.6 & Chap. 7] and the references therein for a thorough review of this topic. To the authors' knowledge, the argument in Lemma 68.12 seems new. □

68.3.3 Error analysis

The error analysis of the Crank–Nicolson scheme is similar to that of the BDF2 scheme. It is done in §70.2.2 in the framework of continuous Petrov–Galerkin time schemes. The error estimate for the Crank–Nicolson scheme is obtained by setting $k := 1$ in Theorem 70.11. Consider the approximation of the heat equation with H^1-conforming finite elements. Under appropriate smoothness assumptions one obtains an error estimate in the $\ell^2(J; H_0^1(D))$-norm that decays as $\mathcal{O}(h^r + \tau^2)$ and an error estimate in the $\ell^\infty(\overline{J}; L^2(D))$-norm decays as $\mathcal{O}(h^{r+s} + \tau^2)$, where $s \in (0,1]$ is the elliptic regularity pickup index ($s = 1$ if there is full elliptic regularity pickup).

Exercises

Exercise 68.1 (Heat equation). Write the error estimates for the heat equation using the BDF2 time discretization in the setting of Remark 68.8.

Exercise 68.2 (Inverse inequality on A_h). Prove (68.22). (*Hint*: observe that $\|A_h(v_h)\|_L = \max_{w_h\in V_h} \frac{|(A_h(v_h),w_h)_L|}{\|w_h\|_L}$ and use the boundedness of a.)

Exercise 68.3 (Discrete Gronwall's lemma). The objective of this exercise is to prove the following discrete Gronwall's lemma. Let $(\gamma_n)_{n\in\mathcal{N}_\tau}$, $(a_n)_{n\in\mathcal{N}_\tau}$, $(b_n)_{n\in\mathcal{N}_\tau}$, $(c_n)_{n\in\mathcal{N}_\tau}$ be sequences of real numbers. Let $B\in\mathbb{R}$. Assume that

$$\gamma_n\in(0,1),\qquad a_n\ge 0,\qquad b_n\ge 0, \tag{68.23a}$$

$$a_n+\sum_{l\in\{1:n\}} b_l \le \sum_{l\in\{1:n\}} \gamma_l a_l + \sum_{l\in\{1:n\}} c_l + B, \tag{68.23b}$$

for all $n\in\mathcal{N}_\tau$. Then we have

$$a_n+\sum_{l\in\{1:n\}} b_l \le \sum_{l\in\{1:n\}} c_l \prod_{\mu\in\{l:n\}} \frac{1}{1-\gamma_\mu} + B\prod_{\mu\in\{1:n\}}\frac{1}{1-\gamma_\mu}. \tag{68.24}$$

(i) Let $d_n\coloneqq\sum_{l\in\{1:n\}}\gamma_l a_l+\sum_{l\in\{1:n\}}(c_l-b_l)+B-a_n$ and let $S_n\coloneqq d_n+a_n+\sum_{l\in\{1:n\}} b_l$. Show that $S_n(1-\gamma_n)\le S_{n-1}+c_n$ for all $n\ge 2$. (*Hint*: observe that $a_n\le S_n$.) (ii) Show by induction that $S_n\le\sum_{l\in\{1:n\}} c_l\prod_{\mu\in\{l:n\}}\frac{1}{1-\gamma_\mu}+B\prod_{\mu\in\{1:n\}}\frac{1}{1-\gamma_\mu}$. Conclude. (*Hint*: (68.23b) means that $d_n\ge 0$.) *Note*: if one replaces the assumption (68.23b) by the assumption $(1+\gamma)a_n\le a_{n-1}+c_n$ which implies (68.23b) with $b_l\coloneqq 0$, $B\coloneqq a_0$, and $\gamma_l\coloneqq-\gamma$ for all $l\in\{1{:}n\}$, the incremental Gronwall lemma from Exercise 67.1 leads to the same bound on a_n as (68.24). The incremental Gronwall lemma only requires that $\gamma>-1$, whereas the discrete Gronwall lemma requires that $\gamma_l\in(0,1)$ (i.e., $\gamma\in(-1,0)$ if one sets $\gamma_l\coloneqq\gamma$).

Exercise 68.4 (Variant on BDF2). The objective of this exercise is to revisit the stability argument for BDF2 proposed in Thomée [273, p. 18]. Consider the setting introduced in §68.2 and the scheme (68.1). (i) Show that for all $k\ge 2$

$$\begin{aligned}(\tfrac{3}{2}u_h^k-2u_h^{k-1}+\tfrac{1}{2}u_h^{k-2},u_h^k)_L &= \|u_h^k\|_L^2-\|u_h^{k-1}\|_L^2-\tfrac{1}{4}(\|u_h^k\|_L^2-\|u_h^{k-2}\|_L^2)\\ &\quad+\|u_h^k-u_h^{k-1}\|_L^2-\tfrac{1}{4}\|u_h^k-u_h^{k-2}\|_L^2.\end{aligned}$$

(ii) Prove that $\sum_{k\in\{2:n\}}\|u_h^k\|_L^2-\|u_h^{k-1}\|_L^2-\frac{1}{4}(\|u_h^k\|_L^2-\|u_h^{k-2}\|_L^2)=\frac{3}{4}\|u_h^n\|_L^2-\frac{1}{4}\|u_h^{n-1}\|_L^2-\frac{3}{4}\|u_h^1\|_L^2+\frac{1}{4}\|u_h^0\|_L^2$, and that

$$\sum_{k\in\{2:n\}}\|u_h^k-u_h^{k-1}\|_L^2-\tfrac{1}{4}\|u_h^k-u_h^{k-2}\|_L^2\ge\tfrac{1}{2}\|u_h^n-u_h^{n-1}\|_L^2-\tfrac{1}{2}\|u_h^1-u_h^0\|_L^2.$$

(iii) Show that

$$\begin{aligned}(u_h^1-u_h^0,u_h^1)_L+\sum_{k\in\{2:n\}}(\tfrac{3}{2}u_h^k-2u_h^{k-1}+\tfrac{1}{2}u_h^{k-2},u_h^k)_L&\\ \ge\tfrac{3}{4}\|u_h^n\|_L^2-\tfrac{1}{4}\|u_h^{n-1}\|_L^2-\tfrac{1}{4}\|u_h^1\|_L^2-\tfrac{1}{4}\|u_h^0\|_L^2.&\end{aligned}$$

(iv) Assuming that $f^k \in L$ for all $k \in \mathcal{N}_\tau$, show that

$$3\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\alpha\|u_h^k\|_V^2 \le \|u_h^0\|_L^2 + \|u_h^1\|_L^2 + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L\|u_h^k\|_L.$$

(v) Letting $m \in \{0{:}n\}$ be the index s.t. $\|u_h^m\|_L := \|u_{h\tau}\|_{\ell^\infty(\overline{J};L)}$, show that

$$2\|u_{h\tau}\|_{\ell^\infty(\overline{J};L)} \le \|u_h^0\|_L + \|u_h^1\|_L + \sum_{k\in\{1:n\}} 4\tau\|f^k\|_L.$$

(vi) Conclude that $\|u_{h\tau}\|_{\ell^\infty(\overline{J};L)} \le \|u_h^0\|_L + \frac{\tau}{2}\|f^1\|_L + \sum_{k\in\{1:n\}} 2\tau\|f^k\|_L$.
(vii) Modify the argument to account for $f^k \in V'$ instead of $f^k \in L$ for all $k \ge 2$, and $f^1 = f^1_{V'} + f^1_L$, where $f^1_{V'} \in V'$ and $f^1_L \in L$, and prove that

$$\|u_{h\tau}\|^2_{\ell^\infty(\overline{J};L^2)} \le \frac{5}{2}\|u_h^0\|_L^2 + 6\tau^2\|f_L^1\|_L^2 + \sum_{k\in\{1:n\}} \frac{\tau}{\alpha}\|\tilde{f}^k\|_{V'}^2.$$

Exercise 68.5 (Variant of Crank–Nicolson scheme). Consider the following variant of the Crank–Nicolson scheme: after setting $u_h^0 := \mathcal{P}_{V_h}(u^0)$, we construct the sequence of functions $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ such that

$$(u_h^n - u_h^{n-1}, w_h)_L + \frac{1}{2}\tau\big(a^n(u_h^n, w_h) + a^{n-1}(u_h^{n-1}, w_h)\big) = \tau\langle \check{f}^{n-\frac{1}{2}}, w_h\rangle_{V',V},$$

for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, with $a^n(\cdot,\cdot) := a(t_n;\cdot,\cdot)$, $a^{n-1}(\cdot,\cdot) := a(t_{n-1};\cdot,\cdot)$, and $\check{f}^{n-\frac{1}{2}} := \frac{1}{2}(f(t_n) + f(t_{n-1})) \in V'$. Assume that $f \in C^0(\overline{J};L)$ and that the restriction (68.20) on the time step holds true. Prove again the bound (68.21) on $\|u_h^n\|_L^2$ with $\check{f}^{k-\frac{1}{2}}$ in lieu of $f^{k-\frac{1}{2}}$ on the right-hand side. (*Hint*: adapt the proof of Lemma 68.12 by starting from the identity $u_h^n + \frac{1}{2}\tau A_h^n(u_h^n) = u_h^{n-1} - \frac{1}{2}\tau A_h^{n-1}(u_h^{n-1}) + \check{f}^{n-\frac{1}{2}}$.) *Note*: deriving an $\ell^2(J;V)$-stability estimate as in Lemma 68.9 is more delicate with this variant of the Crank–Nicolson scheme.

Chapter 69

Discontinuous Galerkin in time

In the previous two chapters, we have used finite differences to approximate the time derivative in the space semi-discrete parabolic problem (66.6). We now adopt a different viewpoint directly relying on the space-time weak formulation from Chapter 65. The time approximation is realized by using piecewise polynomial functions over the time mesh. The test functions are discontinuous at the time nodes, thereby allowing for a time-stepping process, i.e., the discrete formulation decouples into local problems over each time step. This leads to two new families of schemes. In the present chapter, we study the discontinuous Galerkin method in time, where the trial functions are also discontinuous at the time nodes. In the next chapter, we study the continuous Petrov–Galerkin methods where they are continuous. The lowest-order version of the discontinuous Galerkin technique is the implicit Euler scheme, and the lowest-order version of the Petrov–Galerkin technique is the Crank–Nicolson scheme. All these schemes are implicit Runge–Kutta methods.

69.1 Setting for the time discretization

Recall that we divide the time interval $J := (0,T)$, $T > 0$, into N subintervals J_n for all $n \in \mathcal{N}_\tau := \{1{:}N\}$, where N is a positive natural number. To simplify the notation, we assume that all the time intervals are of equal length, i.e., we define the *time step* $\tau := \frac{T}{N}$, the *discrete time nodes* $t_n := n\tau$, for all $n \in \overline{\mathcal{N}}_\tau := \{0{:}N\}$, and we set $J_n := (t_{n-1}, t_n]$ for all $n \in \mathcal{N}_\tau$. Notice that here J_n is open at its left end and closed at its right end. The time mesh is defined as $J_\tau := \bigcup_{n\in\mathcal{N}_\tau} J_n$. For all $n \in \mathcal{N}_\tau$, we define the mapping

$$T_n : \widehat{J} := (-1,1] \to J_n, \qquad T_n(s) := \frac{1}{2}(t_{n-1}+t_n) + \frac{1}{2}\tau s, \ \forall s \in \widehat{J}. \tag{69.1}$$

Let H be a real Hilbert space composed of functions defined on the space domain $D \subset \mathbb{R}^d$. Let $k \geq 0$ be the polynomial degree used for the time approximation of the functions in $L^1(J;H)$. We denote by $\mathbb{P}_k(\widehat{J};\mathbb{R})$ the real vector space composed of the restrictions to $\widehat{J}$ of the polynomials in $\mathbb{P}_k(\mathbb{R};\mathbb{R})$. We adopt a similar definition for $\mathbb{P}_k(J_n;\mathbb{R})$ for all $n \in \mathcal{N}_\tau$, and observe that $p \in \mathbb{P}_k(J_n;\mathbb{R})$ iff $p \circ T_n \in \mathbb{P}_k(\widehat{J};\mathbb{R})$. We define

$$\mathbb{P}_k(J_n;H) := \mathbb{P}_k(J_n;\mathbb{R}) \otimes H, \tag{69.2}$$

i.e., $v \in \mathbb{P}_k(J_n;H)$ if there are $m \in \mathbb{N}$ and $\{(\mathsf{V}_i, p_i) \in H \times \mathbb{P}_k(J_n;\mathbb{R})\}_{i\in\{0:m\}}$ such that $v(t) = \sum_{i\in\{0:m\}} \mathsf{V}_i p_i(t)$. Also, given any basis $\{\psi_l\}_{l\in\{0:k\}}$ of $\mathbb{P}_k(\widehat{J};\mathbb{R})$, $v \in \mathbb{P}_k(J_n;H)$ if there are $\{\mathsf{V}_l \in$

$H\}_{l\in\{0:k\}}$ s.t. $v=\sum_{l\in\{0:k\}}\mathsf{V}_m(\psi_l\circ T_n^{-1})$. Notice that functions in $\mathbb{P}_k(J_n;\mathbb{R})$ are not defined at t_{n-1} since J_n is open at t_{n-1}. The (broken) space composed of the H-valued functions that are piecewise polynomials of degree at most k on the time mesh J_τ is defined as

$$P_k^{\mathrm{b}}(J_\tau;H):=\{v_\tau:(0,T]\to H\mid v_{\tau|J_n}\in\mathbb{P}_k(J_n;H),\ \forall n\in\mathcal{N}_\tau\}. \tag{69.3}$$

The functions in $P_k^{\mathrm{b}}(J_\tau;H)$ are not necessarily continuous at the discrete time nodes, and they are unspecified at $t=0$. If H is finite-dimensional, then $P_k^{\mathrm{b}}(J_\tau;H)$ has dimension $N(k+1)\times\dim(H)$. We also consider the space

$$P_k^{\mathrm{b}}(\overline{J}_\tau;H):=\{v_\tau:\overline{J}:=[0,T]\to H\mid v_{\tau|(0,T]}\in P_k^{\mathrm{b}}(J_\tau;H)\}. \tag{69.4}$$

Hence, every function $v_\tau\in P_k^{\mathrm{b}}(\overline{J}_\tau;H)$ can be represented by the pair $(v_\tau(0),v_{\tau|(0,T]})\in H\times P_k^{\mathrm{b}}(J_\tau;H)$. This means that the space $P_k^{\mathrm{b}}(\overline{J}_\tau;H)$ is isomorphic to $H\times P_k^{\mathrm{b}}(J_\tau;H)$. By definition, every function $v_\tau\in P_k^{\mathrm{b}}(\overline{J}_\tau;H)$ is left-continuous at the discrete time nodes t_n for all $n\in\mathcal{N}_\tau$, i.e., $v(t_n)=v(t_n^-):=\lim_{t\uparrow t_n}v(t)$, and we define the jump of v_τ at the left end of the time interval J_n (i.e., at t_{n-1}) by

$$[\![v_\tau]\!]_{n-1}:=v_\tau(t_{n-1}^+)-v_\tau(t_{n-1}),\qquad v_\tau(t_{n-1}^+):=\lim_{h\downarrow0}v_\tau(t_{n-1}+h). \tag{69.5}$$

The time-discrete setting is illustrated in the left panel of Figure 69.1. Another useful space is the subspace of $P_k^{\mathrm{b}}(\overline{J}_\tau;H)$ composed of the functions $v_\tau:\overline{J}\to H$ that are continuous in time: For all $k\geq1$, we set

$$P_k^{\mathrm{g}}(\overline{J}_\tau;H):=P_k^{\mathrm{b}}(\overline{J}_\tau;H)\cap C^0(\overline{J};H). \tag{69.6}$$

If H is finite-dimensional, then $P_k^{\mathrm{g}}(\overline{J}_\tau;H)$ has dimension $(Nk+1)\times\dim(H)$.

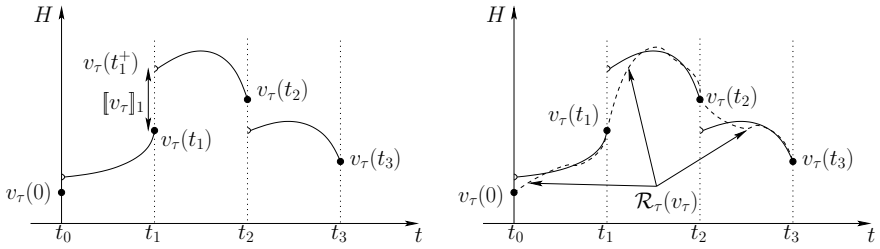

Figure 69.1: Example of time-discrete function $v_\tau\in P_k^{\mathrm{b}}(\overline{J}_\tau;H)$ (left panel) and its time reconstruction $\mathcal{R}_\tau(v_\tau)\in P_{k+1}^{\mathrm{g}}(\overline{J}_\tau;H)$ (right panel, bold dashed curve, see Definition 69.5).

Recall that the model parabolic problem (66.3) is formulated using the Gelfand triple $(V,L\equiv L',V')$, the Hilbert spaces $X:=\{v\in L^2(J;V)\mid\partial_t v\in L^2(J;V')\}$ and $Y:=L\times L^2(J;V)$, and the forms $b:X\times Y\to\mathbb{R}$ and $\ell:Y\to\mathbb{R}$ such that $b(v,y):=(v(0),w)_L+\int_J\big(\langle\partial_t v(t),z(t)\rangle_{V',V}+a(t;v(t),z(t))\big)\mathrm{d}t$ and $\ell(y):=(u_0,w)_L+\int_J\langle f(t),z(t)\rangle_{V',V}\mathrm{d}t$ for all $v\in X$ and $y:=(w,z)\in Y$. In the entire chapter, we assume for simplicity that it is possible to consider pointwise values in time of the bilinear form a. Let $(V_h)_{h\in\mathcal{H}}$ be a sequence of finite-dimensional subspaces of V. Let us set $X_h:=H^1(J;V_h)$, $Y_h:=V_h\times L^2(J;V_h)$. Given $u_0\in L$ and $f\in C^0(\overline{J};V')$ (for simplicity), our starting point is the semi-discrete problem (66.6):

$$\begin{cases}\text{Find } u_h\in X_h \text{ such that}\\ b(u_h,y_h)=\ell(y_h),\quad\forall y_h\in Y_h.\end{cases} \tag{69.7}$$

Recalling that $\mathcal{P}_{V_h} : L \to V_h$ is the L-orthogonal projection from L onto V_h, (i.e., $(z-\mathcal{P}_{V_h}(z), w_h)_L = 0$ for all $z \in L$ and all $w_h \in V_h$), let us set $f_h(t) := \mathcal{P}_{V_h}(f(t)) \in V_h$ for all $t \in J$. Let us also define $A_h(t) : V_h \to V_h$ s.t. $(A_h(t)(v_h), w_h)_L := a_h(t; v_h, w_h)$ for all $v_h, w_h \in V_h$ and all $t \in J$. Then, setting $u_h(0) = \mathcal{P}_{V_h}(u_0)$, the semi-discrete problem (69.7) amounts to finding $u_h \in H^1(J; V_h)$ so that the following holds true for all $t \in J$:

$$\partial_t u_h(t) = f_h(t) - A_h(t)(u_h(t)). \tag{69.8}$$

Remark 69.1 (Method of lines vs. Rothe's method). In the method of lines, one starts with the discretization in space using, e.g., finite elements. This leads to a finite set of coupled ODEs (see(69.8)) which is then discretized in time. An alternative approach consists of applying a time discretization technique to the weak formulation (66.3) first. This leads to a finite set of coupled PDEs which is then discretized in space using, e.g., finite elements. This viewpoint is sometimes called *Rothe's method* in the literature. In many situations, both methods yield the same fully discrete problem. □

69.2 Formulation of the method

Let $k \geq 0$ be the polynomial degree for the time discretization. In this section, we study the dG(k) scheme to approximate in time the semi-discrete problem (69.7). This leads to a nonconforming time approximation. The lowest-order version, dG(0), is the implicit Euler scheme studied in §67.1.

69.2.1 Quadratures and interpolation

Let $\{\xi_l\}_{l\in\{1:k+1\}}$ be the (right-sided) *Gauss–Radau nodes* in the reference interval $\widehat{J} := (-1, 1]$, and let $\{\omega_l\}_{l\in\{1:k+1\}}$ be the corresponding weights. This set of nodes and weights gives a quadrature of order $2k$ (see Proposition 6.7). Using the mapping $T_n : \widehat{J} \to J_n$ defined in (69.1) for all $n \in \mathcal{N}_\tau$, we obtain a quadrature in J_n with $t_{n,l} := T_n(\xi_l)$ and $\omega_{n,l} := \frac{\tau}{2}\omega_l$ for all $l \in \{1:k+1\}$. Notice that setting $c_l := \frac{1+\xi_l}{2} \in (0, 1]$, we have $t_{n,l} = t_{n-1} + c_l\tau$. We introduce the discrete measure $\mu^{\mathrm{GR}}_{k+1}(\mathrm{d}t)$ defined on J so that

$$\int_J g(t)\,\mu^{\mathrm{GR}}_{k+1}(\mathrm{d}t) := \sum_{n\in\mathcal{N}_\tau}\int_{J_n} g(t)\,\mu^{\mathrm{GR}}_{k+1}(\mathrm{d}t) := \sum_{n\in\mathcal{N}_\tau}\sum_{l\in\{1:k+1\}} \omega_{n,l}g(t_{n,l}),$$

for all $g \in C^0(\overline{J}; \mathbb{R})$. We slightly abuse the terminology by using the same symbol for the discrete measure on J and its restriction to the interval J_n.

Let $\mathcal{L}_l(\xi) := \prod_{j\in\{1:k+1\}\setminus\{l\}} \frac{\xi-\xi_j}{\xi_l-\xi_j} \in \mathbb{P}_k(\widehat{J}; \mathbb{R})$ be the Lagrange polynomial based on the Gauss–Radau nodes and associated with the l-th node, i.e., we have $\mathcal{L}_l(\xi_{l'}) = \delta_{ll'}$ for all $l, l' \in \{1:k+1\}$. Let $Z \in \{V', L, V_h\}$ and $\mathcal{I}^{\mathrm{GR}}_k : H^1(J; Z) \to P^{\mathrm{b}}_k(J_\tau; Z)$ be the Lagrange *interpolation operator* associated with the Gauss–Radau nodes, i.e., we set for all $v \in H^1(J; Z) \hookrightarrow C^0(\overline{J}; Z)$,

$$\mathcal{I}^{\mathrm{GR}}_k(v)_{|J_n} := \sum_{l\in\{1:k+1\}} v(t_{n,l})\mathcal{L}_l \circ T_n^{-1}, \qquad \forall n \in \mathcal{N}_\tau. \tag{69.9}$$

Since $\mathcal{I}^{\mathrm{GR}}_k$ is $L^\infty(J; Z)$-stable uniformly w.r.t. τ and leaves $P^{\mathrm{b}}_k(J_\tau; Z)$ pointwise invariant, there is c such that for all τ and all $v \in H^{k+1}(J; Z)$,

$$\|v - \mathcal{I}^{\mathrm{GR}}_k(v)\|_{L^2(J;Z)} \leq c\,\tau^{k+1}|v|_{H^{k+1}(J;Z)}. \tag{69.10}$$

Moreover, we shall use the following identities (see Exercise 69.1): For all $p \in P_k^{\rm b}(J_\tau;L)$ and all $v,w \in H^1(J;L)$,

$$\int_J (v, \mathcal{I}_k^{\rm GR}(w))_L \, \mu_{k+1}^{\rm GR}(\mathrm{d}t) = \int_J (v,w)_L \, \mu_{k+1}^{\rm GR}(\mathrm{d}t), \tag{69.11a}$$

$$\int_J (p, \mathcal{I}_k^{\rm GR}(w))_L \mathrm{d}t = \int_J (p,w)_L \, \mu_{k+1}^{\rm GR}(\mathrm{d}t). \tag{69.11b}$$

69.2.2 Discretization in time

The time-discrete trial and test spaces are defined by

$$X_{h\tau} := P_k^{\rm b}(\overline{J}_\tau; V_h), \qquad Y_{h\tau} := X_{h\tau}. \tag{69.12}$$

We then consider the bilinear form b_τ such that for all $(v_{h\tau}, y_{h\tau}) \in X_{h\tau}\times Y_{h\tau}$,

$$\begin{aligned} b_\tau(v_{h\tau}, y_{h\tau}) &:= (v_{h\tau}(0), y_{h\tau}(0))_L + \sum_{n\in\mathcal{N}_\tau} \int_{J_n} (\partial_t v_{h\tau}(t), y_{h\tau}(t))_L \mu_{k+1}^{\rm GR}(\mathrm{d}t) \\ &+ \sum_{n\in\mathcal{N}_\tau} (\llbracket v_{h\tau}\rrbracket_{n-1}, y_{h\tau}(t_{n-1}^+))_L + \int_J a(t; v_{h\tau}(t), y_{h\tau}(t)) \, \mu_{k+1}^{\rm GR}(\mathrm{d}t). \end{aligned} \tag{69.13}$$

Similarly, we consider the linear form ℓ_τ such that for all $y_{h\tau} \in Y_{h\tau}$,

$$\ell_\tau(y_{h\tau}) := (u_0, y_{h\tau}(0))_L + \int_J \langle f(t), y_{h\tau}(t)\rangle_{V',V} \, \mu_{k+1}^{\rm GR}(\mathrm{d}t).$$

We observe that for all $n \in \mathcal{N}_\tau$,

$$\int_{J_n} (\partial_t v_{h\tau}(t), y_{h\tau}(t))_L \mu_{k+1}^{\rm GR}(\mathrm{d}t) = \int_{J_n} (\partial_t v_{h\tau}(t), y_{h\tau}(t))_L \mathrm{d}t, \tag{69.14}$$

since the integrand is in $\mathbb{P}_{2k-1}(J_n;\mathbb{R}) \subset \mathbb{P}_{2k}(J_n;\mathbb{R})$ and the quadrature is of order $2k$. The same remark applies to the integral $\int_{J_n} a(t; v_{h\tau}(t), y_{h\tau}(t))_L \mathrm{d}t$ if the bilinear form a is time-independent. The dG(k) scheme consists of solving the following space-time discrete problem:

$$\begin{cases} \text{Find } u_{h\tau} \in X_{h\tau} \text{ such that} \\ b_\tau(u_{h\tau}, y_{h\tau}) = \ell_\tau(y_{h\tau}), \quad \forall y_{h\tau} \in Y_{h\tau}. \end{cases} \tag{69.15}$$

Notice that (69.15) is a square linear system of size $\dim(X_{h\tau}) = \dim(Y_{h\tau}) = (1+N(k+1)) \times \dim(V_h)$. The time approximation is *nonconforming* since the trial functions can jump at the discrete time nodes. Hence, $\partial_t u_{h\tau}$ is not necessarily integrable in time over J, but it is integrable over all the time intervals J_n, $n \in \mathcal{N}_\tau$. To account for these discontinuities, the time integral of the time derivative has been transformed into a sum over the time intervals $\{J_n\}_{n\in\mathcal{N}_\tau}$, and the corresponding jump terms have been added in (69.13).

Proposition 69.2 (Localization). *The* dG(k) *solution $u_{h\tau}$ (if it exists) is such that $u_{h\tau}(0) = \mathcal{P}_{V_h}(u_0)$ and for all $q \in \mathbb{P}_k(J_n;V_h)$ and all $n \in \mathcal{N}_\tau$,*

$$\begin{aligned} &\int_{J_n} (\partial_t u_{h\tau}(t), q(t))_L \mathrm{d}t + (\llbracket u_{h\tau}\rrbracket_{n-1}, q(t_{n-1}^+))_L \\ &+ \int_{J_n} a(t; u_{h\tau}(t), q(t)) \, \mu_{k+1}^{\rm GR}(\mathrm{d}t) = \int_{J_n} \langle f(t), q(t)\rangle_{V',V} \, \mu_{k+1}^{\rm GR}(\mathrm{d}t). \end{aligned} \tag{69.16}$$

Proof. The assertion on $u_{h\tau}(0)$ follows by considering the test function $y_{h\tau} \in Y_{h\tau}$ with $y_{h\tau}(0) := w_h$ arbitrary in V_h and $y_{h\tau|(0,T]} := 0$. The identity (69.16) follows by taking the test function $y_{h\tau} \in Y_{h\tau}$ with $y_{h\tau}(0) := 0$, $y_{h\tau|J_n} := q$ arbitrary in $\mathbb{P}_k(J_n; V_h)$, and $y_{h\tau|(0,T]\setminus J_n} := 0$ for all $n \in \mathcal{N}_\tau$. □

Proposition 69.2 shows that the dG(k) scheme leads to a time-stepping procedure, where $u_{h\tau}(0)$ is computed first as $u_{h\tau}(0) := \mathcal{P}_{V_h}(u_0)$, and then the restrictions $u_{h\tau|J_n}$ are computed sequentially by solving (69.16) for $n = 1, 2, \ldots, N$. Notice that the value $u_{h\tau}(t_{n-1})$ from the previous time interval (or the initial condition if $n = 1$) is needed to compute $[\![u_{h\tau}]\!]_{n-1}$ as defined in (69.5).

Example 69.3 (Implicit Euler, dG(0)). Let us take $k := 0$. Then $\partial_t u_{h\tau|J_n} = 0$ and $[\![u_{h\tau}]\!]_{n-1} = u_h(t_n) - u_h(t_{n-1})$ for all $n \in \mathcal{N}_\tau$. Since the test function q in (69.16) is constant in time and since for $k = 0$, the only (right-sided) Gauss–Radau node in J_n is $t_{n,1} := t_n$, we obtain $(u_h(t_n) - u_h(t_{n-1}), w_h)_L + \tau a(t_n, u_h(t_n), w_h) = \tau \langle f(t_n), w_h\rangle_{V',V}$ for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$. Thus, we recover the implicit Euler scheme studied in §67.1. Notice that $u_{h\tau|J_n} = u(t_n)$ for all $n \in \mathcal{N}_\tau$. □

Remark 69.4 (Literature). Discontinuous Galerkin methods in time have been originally considered by Hulme [192], Lesaint and Raviart [215], Jamet [196], Delfour et al. [102], Johnson et al. [201], Eriksson et al. [116]. We refer the reader to Schötzau and Schwab [248], Akrivis and Makridakis [7], Chrysafinos and Walkington [87], Schötzau and Wihler [249], Thomée [273, Chap. 12], Schmutz and Wihler [247] for further results on the analysis of dG(k) methods for parabolic problems. □

69.2.3 Reformulation using a time reconstruction operator

A useful reformulation of (69.16) consists of combining together the time derivative and the jump terms by means of a suitable time reconstruction operator in the same spirit as the discrete gradients introduced in §38.4 in the context of discontinuous Galerkin methods in space.

Definition 69.5 (Time reconstruction). *The time reconstruction operator*

$$\mathcal{R}_\tau : X_{h\tau} := P_k^{\mathrm{b}}(\overline{J}_\tau; V_h) \to P_{k+1}^{\mathrm{g}}(\overline{J}_\tau; V_h)$$

is defined by setting for all $v_{h\tau} \in X_{h\tau}$ and all $n \in \mathcal{N}_\tau$,

$$\mathcal{R}_\tau(v_{h\tau})(t_{n-1}) := v_{h\tau}(t_{n-1}), \tag{69.17a}$$

$$\mathcal{R}_\tau(v_{h\tau})(t_{n,l}) := v_{h\tau}(t_{n,l}), \qquad \forall l \in \{1{:}k+1\}. \tag{69.17b}$$

This definition makes sense since over each interval $\overline{J}_n$, $\mathcal{R}_\tau(v_{h\tau})$ is the Lagrange interpolation of $v_{h\tau}$ at the Lagrange nodes $\{t_{n-1}, \{t_{n,l}\}_{l\in\{1:k+1\}}\}$. The time reconstruction operator is illustrated in the right panel of Figure 69.1. The key property of $\mathcal{R}_\tau$ we are going to use is the following.

Lemma 69.6 (Derivative of $\mathcal{R}_\tau$). *The following holds true for all $v_{h\tau} \in X_{h\tau}$, all $q \in \mathbb{P}_k(J_n; V_h)$, and all $n \in \mathcal{N}_\tau$,*

$$\int_{J_n} (\partial_t(\mathcal{R}_\tau(v_{h\tau})), q)_L \mathrm{d}t = \int_{J_n} (\partial_t v_{h\tau}, q)_L \mathrm{d}t + ([\![v_{h\tau}]\!]_{n-1}, q(t_{n-1}^+))_L. \tag{69.18}$$

Proof. Recalling that the (right-sided) Gauss–Radau rule is of order $2k$, using (69.17), and $(\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau}, \partial_t q)_L \in \mathbb{P}_{2k}(J_n; \mathbb{R})$, we obtain the identity $\int_{J_n} (\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau}, \partial_t q)_L \mathrm{d}t = 0$. Then an integration by parts gives

$$\begin{aligned}\int_{J_n} (\partial_t(\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau}), q)_L \mathrm{d}t &= [(\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau}, q)_L]_{t_{n-1}^+}^{t_n} \\ &= ([\![v_{h\tau}]\!]_{n-1}, q(t_{n-1}^+))_L,\end{aligned}$$

since $(\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau})(t_n) = 0$ and $(\mathcal{R}_\tau(v_{h\tau}) - v_{h\tau})(t_{n-1}^+) = -[\![v_{h\tau}]\!]_{n-1}$. □

Recalling that $\mathcal{P}_{V_h} : L \to V_h$ is the L-orthogonal projection from L onto V_h, we set $f_h(t) := \mathcal{P}_{V_h}(f(t)) \in V_h$ for all $t \in J$. We also define $A_h(t) : V_h \to V_h$ s.t. $(A_h(t)(v_h), w_h)_L := a_h(t; v_h, w_h)$ for all $v_h, w_h \in V_h$ and all $t \in J$.

Proposition 69.7 (Reformulations). (i) (69.16) *is equivalent to*

$$\begin{aligned}&\int_{J_n} (\partial_t \mathcal{R}_\tau(u_{h\tau})(t), q(t))_L \mathrm{d}t + \int_{J_n} a(t; u_{h\tau}(t), q(t))\, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t)\\ &= \int_{J_n} \langle f(t), q(t)\rangle_{V',V}\, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t), \quad \forall q \in \mathbb{P}_k(J_n; V_h),\ \forall n \in \mathcal{N}_\tau. \end{aligned} \tag{69.19}$$

(ii) (69.19) *is equivalent to the following equations: For all* $l \in \{1\!:\!k+1\}$ *and all* $n \in \mathcal{N}_\tau$,

$$\partial_t \mathcal{R}_\tau(u_{h\tau})(t_{n,l}) + A_h(t_{n,l})(u_{h\tau}(t_{n,l})) = f_h(t_{n,l}). \tag{69.20}$$

Proof. The equivalence of (69.16) and (69.19) is a direct consequence of (69.18). To prove the equivalence of (69.16) and (69.20), we observe that, since the Gauss–Radau quadrature is of order $2k$, (69.19) can be rewritten as follows: For all $q \in \mathbb{P}_k(J_n; \mathbb{R})$,

$$\sum_{l \in \{1:k+1\}} \omega_{n,l} q(t_{n,l}) \Big(\partial_t \mathcal{R}_\tau(u_{h\tau})(t_{n,l}) + A_h(t_{n,l})(u_h^{n,l}) - f_h(t_{n,l}), v_h \Big)_L = 0.$$

Using $\{\mathcal{L}_l\}_{l \in \{1:k+1\}}$ as test functions yields the assertion. □

Example 69.8 ($k = 0$). Consider the implicit Euler scheme dG(0). The linear Lagrange interpolant of $v_{h\tau}$ over $\overline{J}_n$ using the time nodes t_{n-1} and t_n is $\mathcal{R}_\tau(v_{h\tau})(t) = \frac{t_n - t}{\tau} v_h(t_{n-1}) + \frac{t - t_{n-1}}{\tau} v_h(t_n)$ for all $t \in J_n$ and all $n \in \mathcal{N}_\tau$. (Recall that $v_{h\tau|J_n} = v_h(t_n)$ for all $n \in \mathcal{N}_\tau$ since $v_{h\tau}$ is piecewise constant in time.) □

Remark 69.9 (Other definition). Letting $\theta_{k+1}(s) := \prod_{l \in \{1:k+1\}} \frac{\xi_l - s}{\xi_l + 1} \in \mathbb{P}_{k+1}(\widehat{J}; \mathbb{R})$, an equivalent definition of $\mathcal{R}_\tau$ is $\mathcal{R}_\tau(v_{h\tau})(0) := v_{h\tau}(0)$ and $\mathcal{R}_\tau(v_{h\tau})_{|J_n} := v_{h\tau|J_n} - [\![v_{h\tau}]\!]_{n-1} \theta_{k+1} \circ T_n^{-1}$ for all $n \in \mathcal{N}_\tau$. Moreover, as shown in Smears [262], one also has $\theta_{k+1} := \frac{(-1)^k}{2}\big(L_k - L_{k+1}\big)$, where $L_m \in \mathbb{P}_m(\widehat{J}; \mathbb{R})$ is the m-th Legendre polynomial (see §6.1). We refer the reader to Exercise 69.4 for the proofs. □

Remark 69.10 (Literature). The operator $\mathcal{R}_\tau$ was introduced in Makridakis and Nochetto [223, Lem. 2.1] and used, e.g., in Schötzau and Wihler [249], Ern and Schieweck [122], Ern et al. [125], Holm and Wihler [185]. □

69.2.4 Equivalence with Radau IIA IRK

It turns out that the dG(k) scheme (69.16) (or (69.19)) is related to an *implicit Runge–Kutta* (IRK) scheme known in the literature as *Radau IIA* (see Makridakis and Nochetto [223, §2.3]). More precisely, let $s \geq 1$ be some integer. An s-stage IRK scheme for solving (69.8) is defined by a set of coefficients, $\{a_{ij}\}_{i,j \in \{1:s\}}$, $\{b_i\}_{i \in \{1:s\}}$, $\{c_i\}_{i \in \{1:s\}}$, and is represented in the literature by its *Butcher tableau* as follows:

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{69.21}$$

Using the Butcher tableau, the time stepping for the semi-discrete problem (69.8) is done as follows (see Hairer and Wanner [174, §2.2], [175, §IV.5], [176, §II.7]): One first sets $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then for all $n \in \mathcal{N}_\tau$, using the abbreviation $t_{n,j} := t_{n-1} + c_j\tau$ for all $j \in \{1\!:\!s\}$, one seeks $\{u_h^{n,i}\}_{i\in\{1:s\}} \subset V_h$ solving the following system of coupled equations:

$$u_h^{n,i} - u_h^{n-1} = \tau \sum_{j\in\{1:s\}} a_{ij}\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big), \tag{69.22}$$

and one sets $u_h^n := u_h^{n-1} + \tau\sum_{j\in\{1:s\}} b_j\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big)$. The s-stage Radau IIA method is defined by setting (see [174, §3.3])

$$a_{ij} := \frac{1}{2}\int_{-1}^{\xi_i} \mathcal{L}_j(\xi)\,\mathrm{d}\xi, \qquad b_i := \frac{1}{2}\int_{-1}^{1} \mathcal{L}_i(\xi)\,\mathrm{d}\xi, \qquad c_i := \frac{\xi_i+1}{2}, \tag{69.23}$$

for all $i,j \in \{1\!:\!s\}$, where $\{\xi_i\}_{i\in\{1:s\}}$ are the right-sided Gauss–Radau quadrature points in $\overline{J}$ and $\mathcal{L}_i \in \mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$ is the Lagrange polynomial associated with the i-th node. Notice that here we have $b_j = a_{sj}$ for all $j \in \{1\!:\!s\}$, which means that $u_h^n = u_h^{n,s}$ (recall that $\xi_s = 1$). Notice also that $t_{n-1} + c_j\tau = T_n(\xi_j)$ for all $j \in \{1\!:\!s\}$ so that the above notation for $t_{n,j}$ is consistent with that used for the Gauss–Radau points in J_n. The Butcher tableaux of the one-stage (implicit Euler), the two-stage, and the three-stage Radau IIA IRK schemes are as follows:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \qquad \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array} \tag{69.24}$$

The following result is proved in Makridakis and Nochetto [223, Lem. 2.3].

Lemma 69.11 (dG(k) ⇔ Radau IIA IRK). *Let $k \ge 0$. Let $u_{h\tau} \in X_{h\tau} := P_k^{\mathrm{b}}(\overline{J}_\tau; V_h)$ and set $\{u_h^{n,l} := u_{h\tau}(t_{n,l})\}_{l\in\{1:k+1\}}$ for all $n \in \mathcal{N}_\tau$. Then $u_{h\tau}$ solves* (69.15) *iff $\{u_h^{n,l}\}_{l\in\{1:s\}}$ solves* (69.22) *with $s := k+1$ for all $n \in \mathcal{N}_\tau$.*

Proof. Assume that $u_{h\tau}$ solves (69.15), i.e., (69.20) by Proposition 69.7. Since $\partial_t\mathcal{R}_\tau(u_{h\tau})_{|J_n} \in \mathbb{P}_k(J_n;V_h)$, (69.20) implies that for all $n \in \mathcal{N}_\tau$,

$$\partial_t\mathcal{R}_\tau(u_{h\tau})_{|J_n} = \sum_{j\in\{1:k+1\}} \big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big)\mathcal{L}_j \circ T_n^{-1}.$$

Integrating this identity over $(t_{n-1}, t_{n,i})$ for all $i \in \{1\!:\!k+1\}$, using the definition of a_{ij} in (69.23), and since $t_{n,i} = T_n(\xi_i)$, this gives

$$u_h^{n,i} - u_h^{n-1} = \tau \sum_{j\in\{1:k+1\}} a_{ij}\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big),$$

since $\mathcal{R}_\tau(u_{h\tau})(t_{n-1}) = u_{h\tau}(t_{n-1}) =: u_h^{n-1}$ and $\mathcal{R}_\tau(u_{h\tau})(t_{n,i}) = u_{h\tau}(t_{n,i}) =: u_h^{n,i}$ (see (69.17)). This shows that $\{u_h^{n,i}\}_{i\in\{1:k+1\}}$ solves (69.22) with $s := k+1$ for all $n \in \mathcal{N}_\tau$. The converse assertion is shown in Exercise 69.2. □

Remark 69.12 (Collocation). In view of (69.20), we say that the dG(k) scheme (or the Radau IIA IRK scheme) is a $(k+1)$-point collocation method using the Gauss–Radau points $\{t_{n,l}\}_{l\in\{1:k+1\}}$ for all $n \in \mathcal{N}_\tau$ (the precise meaning of this assertion is clarified in §70.1.4, see Definition 70.6). □

Remark 69.13 (Final stage). For any s-stage IRK scheme, the update u_h^n is given by $u_h^n = \alpha_0 u_h^{n-1} + \sum_{p\in\{1:s\}} \alpha_p u_h^{n,p}$, where $\alpha_p := \sum_{q\in\{1:s\}} b_q (a^{-1})_{qp}$, $\alpha_0 := 1 - \sum_{p\in\{1:s\}} \alpha_p$, and $(a^{-1})_{pq}$ are the coefficients of the inverse of the Butcher matrix $(a_{pq})_{p,q\in\{1:s\}}$. For the Radau IIA IRK scheme, we have $\alpha_p = 0$ for all $p \in \{0:s-1\}$ and $\alpha_s = 1$. See Exercise 69.6. □

Remark 69.14 (Order conditions). The coefficients of any RK scheme must satisfy some order conditions for the scheme to be of order p; see Theorem 78.5 and Exercise 70.3. In particular, it is necessary to have $\sum_{j\in\{1:s\}} b_j = 1$ to get first-order convergence at least. □

69.3 Stability and error analysis

In this section, we study the stability and the convergence properties of the dG(k) scheme (69.15).

69.3.1 Stability

The key stability mechanism we are going to invoke is a coercivity property of the bilinear form b_τ defined in (69.13). Let $\alpha > 0$ be the coercivity constant of the bilinear form a. We equip $X_{h\tau} := P_k^{\mathrm{b}}(\overline{J}_\tau; V_h)$ with the following norm:

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \|v_{h\tau}\|_{L^2(J;V)}^2 + \frac{1}{2\alpha}\Big(\|v_{h\tau}(T)\|_L^2 + \|v_{h\tau}(0)\|_L^2 + \sum_{n\in\mathcal{N}_\tau} \|[\![v_{h\tau}]\!]_{n-1}\|_L^2\Big).$$

Lemma 69.15 (Coercivity). (i) *The following holds true:*

$$b_\tau(v_{h\tau}, v_{h\tau}) \ge \alpha \|v_{h\tau}\|_{X_{h\tau}}^2, \qquad \forall v_{h\tau} \in X_{h\tau}. \tag{69.25}$$

(ii) *The discrete problem* (69.15) *is well-posed.*

Proof. (i) Let $v_{h\tau} \in X_{h\tau}$. Using the coercivity of $a(t;\cdot,\cdot)$ at the $(k+1)$ Gauss–Radau nodes and the positivity of the weights ω_l, we obtain

$$\begin{aligned}\int_{J_n} a(t; v_{h\tau}(t), v_{h\tau}(t))\, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) &= \sum_{l\in\{1:k+1\}} \omega_{n,l} a(t_{n,l}; v_{h\tau}(t_{n,l}), v_{h\tau}(t_{n,l})) \\ &\ge \alpha \sum_{l\in\{1:k+1\}} \omega_{n,l} \|v_{h\tau}(t_{n,l})\|_V^2 = \alpha \int_{J_n} \|v_{h\tau}(t)\|_V^2 \mathrm{d}t,\end{aligned}$$

since the integrand is in $\mathbb{P}_{2k}(J_n;\mathbb{R})$ and the quadrature order is $2k$. Moreover, using that $\frac{\mathrm{d}}{\mathrm{d}t}\|v_{h\tau}\|_L^2 = 2(\partial_t v_{h\tau}, v_{h\tau})_L$, the identity $2(a-b)a = a^2 - b^2 + (a-b)^2$ with $a := v_{h\tau}(t_{n-1}^+)$ and $b := v_{h\tau}(t_{n-1})$, and the definition (69.5) of $[\![v_{h\tau}]\!]_{n-1}$ gives

$$\begin{aligned}&\sum_{n\in\mathcal{N}_\tau} \Big(\int_{J_n} (\partial_t v_{h\tau}(t), v_{h\tau}(t))_L \mathrm{d}t + ([\![v_{h\tau}]\!]_{n-1}, v_{h\tau}(t_{n-1}^+))_L \Big) \\ &= \frac{1}{2} \sum_{n\in\mathcal{N}_\tau} \Big(\|v_{h\tau}(t_n)\|_L^2 - \|v_{h\tau}(t_{n-1}^+)\|_L^2 + 2([\![v_{h\tau}]\!]_{n-1}, v_{h\tau}(t_{n-1}^+))_L \Big) \\ &= \frac{1}{2} \sum_{n\in\mathcal{N}_\tau} \Big(\|v_{h\tau}(t_n)\|_L^2 - \|v_{h\tau}(t_{n-1})\|_L^2 + \|[\![v_{h\tau}]\!]_{n-1}\|_L^2 \Big) \\ &= \frac{1}{2}\|v_{h\tau}(T)\|_L^2 - \frac{1}{2}\|v_{h\tau}(0)\|_L^2 + \frac{1}{2} \sum_{n\in\mathcal{N}_\tau} \|[\![v_{h\tau}]\!]_{n-1}\|_L^2.\end{aligned}$$

Combining this identity with the lower bound on a proves (69.25).
(ii) The well-posedness of (69.15) results from the Lax–Milgram lemma. □

69.3.2 Error analysis

Let $Z \in \{L, V_h\}$ and recall that $H^1(J;Z) \hookrightarrow C^0(\overline{J};Z)$. To handle the consistency error optimally, we introduce the operator $\Pi_n^k : H^1(J_n;Z) \to \mathbb{P}_k(J_n;Z)$ for all $n \in \mathcal{N}_\tau$ s.t. for all $v \in H^1(J_n;Z)$,

$$\Pi_n^k(v)(t_n) = v(t_n), \tag{69.26a}$$

$$\int_{J_n} (\Pi_n^k(v) - v, q)_L \mathrm{d}t = 0, \qquad \forall q \in \mathbb{P}_{k-1}(J_n;Z). \tag{69.26b}$$

Notice that the statement (69.26b) is void if $k = 0$. The above definition can be extended to $Z := V'$ by replacing the L-inner product in (69.26b) by the duality bracket between V' and V and by taking $q \in \mathbb{P}_{k-1}(J_n;V)$.

We then define $\Pi_\tau^k : H^1(J;Z) \to P_k^{\mathrm{b}}(\overline{J}_\tau;Z)$ by setting $\Pi_\tau^k(v)(0) := v(0)$ and $\Pi_\tau^k(v)_{|J_n} := \Pi_n^k(v_{|J_n})$ for all $n \in \mathcal{N}_\tau$. Since Π_τ^k leaves $\mathbb{P}_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ pointwise invariant and is $L^\infty(J;Z)$-stable uniformly w.r.t. τ, there is c s.t. for all $\tau > 0$ and all $v \in W^{k+1,\infty}(J;Z)$,

$$\|v - \Pi_\tau^k(v)\|_{L^\infty(J;Z)} \le c\,\tau^{k+1} |v|_{W^{k+1,\infty}(J;Z)}. \tag{69.27}$$

The definition of Π_τ^k is motivated by the following result.

Lemma 69.16 (Orthogonality). *The following identity holds true for all $v \in H^1(J;L)$, all $y_\tau \in \mathbb{P}_k^{\mathrm{b}}(\overline{J}_\tau;L)$, and all $n \in \mathcal{N}_\tau$:*

$$\int_{J_n} (\partial_t(v - \Pi_\tau^k(v)), y_\tau)_L \mathrm{d}t - (\llbracket \Pi_\tau^k(v) \rrbracket_{n-1}, y_\tau(t_{n-1}^+))_L = 0. \tag{69.28}$$

Proof. Let $\delta := v - \Pi_\tau^k(v)$. We integrate by parts in time and use the definition (69.26) of Π_n^k and the fact that $\partial_t y_{\tau|J_n} \in \mathbb{P}_{k-1}(J_n;L)$ for all $n \in \mathcal{N}_\tau$. Recalling that $\delta(t_n) = 0$, this yields

$$\int_{J_n} (\partial_t \delta, y_\tau)_L \mathrm{d}t = -\int_{J_n} (\delta, \partial_t y_\tau)_L + \big[(\delta, y_\tau)_L\big]_{t_{n-1}^+}^{t_n} = -(\delta(t_{n-1}^+), y_\tau(t_{n-1}^+))_L.$$

Then (69.28) follows from $\llbracket \Pi_\tau^k(v) \rrbracket_{n-1} = -\delta(t_{n-1}^+)$ since $\Pi_\tau^k(v)(t_{n-1}) = v(t_{n-1}) = v(t_{n-1}^+)$ (recall that $v \in C^0(\overline{J};L)$). □

Remark 69.17 (Other definition). Let L_k be the k-th Legendre polynomial and let $\Xi_{k-1}^n : L^2(J_n;L) \to \mathbb{P}_{k-1}(J_n;L)$ be the $L^2(J_n;L)$-orthogonal projection. Then an equivalent definition of Π_n^k is to set $\Pi_n^k(v)(t) := (v(t_n) - \Xi_{k-1}^n(v)(t_n))L_k(t) + \Xi_{k-1}^n(v)(t)$ for all $v \in H^1(J_n;L)$, all $t \in J_n$, and all $n \in \mathcal{N}_\tau$. □

To separate the time approximation and the space approximation, we assume that we have at hand a time-independent space approximation operator $\Pi_h \in \mathcal{L}(V;V_h)$ with $\|\Pi_h\|_{\mathcal{L}(V;V_h)}$ uniformly bounded w.r.t. $h \in \mathcal{H}$. We could take, for instance, the quasi-interpolation operator $\mathcal{I}_h^{\mathrm{av}}$ or the elliptic projection Π_h^{E} if the bilinear form a is time-independent. We extend Π_h to $\mathcal{L}(L^2(J;V);L^2(J;V_h))$ by setting $\Pi_h(v)(t) := \Pi_h(v(t))$ for all $v \in L^2(J;V)$. Notice that $\partial_t \Pi_h(v) = \Pi_h(\partial_t v)$ for all $v \in H^1(J;V)$ owing to Lemma 64.34. We are also going to use the commuting property $\Pi_\tau^k(\Pi_h(v)) = \Pi_h(\Pi_\tau^k(v))$ for all $v \in H^1(J;V)$; see Exercise 69.7.

We can now state the main result of this section. We extend the $\|\cdot\|_{X_{h\tau}}$-norm defined above to $H^1(J;V) + X_{h\tau}$ (we use the same notation for simplicity). Recalling the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, we set $\rho_\sharp := \max(\rho, T)$.

Theorem 69.18 ($L^2(J;V)$-estimate). *Let $u \in X$ solve (66.3) and $u_{h\tau} \in X_{h\tau}$ solve (69.15). Assume that $u \in H^{k+2}(J;V') \cap W^{k+1,\infty}(J;V)$. Let $c_1(u) := \frac{1}{\alpha}|u|_{H^{k+2}(J;V')} + \xi_\kappa T^{\frac{1}{2}}|u|_{W^{k+1,\infty}(J;V)}$ with the contrast factor $\xi_\kappa := \frac{M}{\alpha}$. There is c s.t. for all $h \in \mathcal{H}$, α, M, and T,*

$$\|u - u_h\|_{X_{h\tau}} \le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0 - \Pi_h(u_0)\|_L + c\Big(\tau^{k+1}c_1(u) + \frac{1}{\alpha}\|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + \xi_\kappa \rho_\sharp^{\frac{1}{2}}\|u - \Pi_h(u)\|_{L^\infty(J;V)}\Big). \tag{69.29}$$

Proof. Let $y_{h\tau} \in X_{h\tau}$. Owing to the regularity assumption on u, we have $(\partial_t u(t_{n,l}), y_{h\tau})_L + a(t_{n,l}; u(t_{n,l}), y_{h\tau}) = \langle f(t_{n,l}), y_{h\tau}\rangle_{V',V}$ for all $n \in \mathcal{N}_\tau$ and all $l \in \{1{:}k{+}1\}$. This gives

$$\begin{aligned} b_\tau(u_{h\tau}, y_{h\tau}) &= (u_0, y_{h\tau}(0))_L + \int_J \langle f, y_{h\tau}\rangle_{V',V}\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\ &= (u_0, y_{h\tau}(0))_L + R(y_{h\tau}) + \int_J (\partial_t u, y_{h\tau})_L \mathrm{d}t + \int_J a(t; u, y_{h\tau})\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t), \end{aligned}$$

with $R(y_{h\tau}) := \int_J (\partial_t u, y_{h\tau})_L \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) - \int_J (\partial_t u, y_{h\tau})_L \mathrm{d}t$. Let us introduce $v_{h\tau} := \Pi_\tau^k(\Pi_h(u)) \in X_{h\tau}$ and let us set $e_{h\tau} := u_{h\tau} - v_{h\tau}$ and $\eta := u - v_{h\tau}$. Using the above calculation for $b_\tau(u_{h\tau}, y_{h\tau})$, we obtain

$$\begin{aligned} b_\tau(u_{h\tau} - v_{h\tau}, y_{h\tau}) &= (u_0 - v_{h\tau}(0), y_{h\tau}(0))_L + R(y_{h\tau}) \\ &\quad + \sum_{n\in\mathcal{N}_\tau}\int_{J_n} (\partial_t(u - v_{h\tau}), y_{h\tau})_L \mathrm{d}t - (\llbracket v_{h\tau}\rrbracket_{n-1}, y_{h\tau}(t_{n-1}^+))_L \\ &\quad + \int_J a(t; u - v_{h\tau}, y_{h\tau})\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t). \end{aligned}$$

By definition of Π_τ^k, we have $v_{h\tau}(0) = \Pi_\tau^k(\Pi_h(u))(0) = \Pi_h(u)(0) = \Pi_h(u_0)$ (notice that $u_0 = u(0) \in V$ since $u \in W^{k+1,\infty}(J;V)$). Moreover, using Lemma 69.16 for the function $v := \Pi_h(u)$ and since $\partial_t(\Pi_h(u)) = \Pi_h(\partial_t u)$ (recall that $u \in H^1(J;V)$ by assumption), we infer that for all $n \in \mathcal{N}_\tau$,

$$\begin{aligned} &\int_{J_n} (\partial_t v_{h\tau}, y_{h\tau})_L \mathrm{d}t + (\llbracket v_{h\tau}\rrbracket_{n-1}, y_{h\tau}(t_{n-1}^+))_L \\ &= \int_{J_n} (\partial_t(\Pi_\tau^k(\Pi_h(u))), y_{h\tau})_L \mathrm{d}t + (\llbracket \Pi_\tau^k(\Pi_h(u))\rrbracket_{n-1}, y_{h\tau}(t_{n-1}^+))_L \\ &= \int_{J_n} (\partial_t(\Pi_h(u)), y_{h\tau})_L \mathrm{d}t = \int_{J_n} (\Pi_h(\partial_t u), y_{h\tau})_L \mathrm{d}t. \end{aligned} \tag{69.30}$$

Hence, we have

$$\begin{aligned} b_\tau(e_{h\tau}, y_{h\tau}) &= (u_0 - \Pi_h(u_0), y_{h\tau}(0))_L + R(y_{h\tau}) \\ &\quad + \int_J (\partial_t u - \Pi_h(\partial_t u), y_{h\tau})_L \mathrm{d}t + \int_J a(t; \eta, y_{h\tau})\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t). \end{aligned}$$

Let $\mathfrak{T}_1, \dots, \mathfrak{T}_4$ denote the four terms on the right-hand side. We have

$$|\mathfrak{T}_1| \le \|u_0 - \Pi_h(u_0)\|_L \|y_{h\tau}(0)\|_L.$$

Furthermore, since $\int_J (\partial_t u, y_{h\tau})_L \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) = \int_J (\mathcal{I}_k^{\mathrm{GR}}(\partial_t u), y_{h\tau})_L \mathrm{d}t$ by (69.11b) with $\mathcal{I}_k^{\mathrm{GR}}$ defined in (69.9), we have $R(y_{h\tau}) = \int_J (\mathcal{I}_k^{\mathrm{GR}}(\partial_t u) - \partial_t u, y_{h\tau})_L \mathrm{d}t$. Using the approximation property (69.10)

of $\mathcal{I}_k^{\mathrm{GR}}$ with $Z := V'$, this gives

$$\begin{aligned}|\mathfrak{T}_2| &\le \|\mathcal{I}_k^{\mathrm{GR}}(\partial_t u) - \partial_t u\|_{L^2(J;V')}\|y_{h\tau}\|_{L^2(J;V)}\\ &\le c\,\tau^{k+1}|u|_{H^{k+2}(J;V')}\|y_{h\tau}\|_{L^2(J;V)}.\end{aligned}$$

Moreover, we have $|\mathfrak{T}_3| \le \|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')}\|y_{h\tau}\|_{L^2(J;V)}$. Finally, since Π_τ^k and Π_h commute, the stability of Π_h and the approximation property (69.27) of Π_τ^k imply that for all $l \in \{1{:}k{+}1\}$ and all $n \in \mathcal{N}_\tau$,

$$\begin{aligned}\|\eta(t_{n,l})\|_V &\le \|(u - \Pi_h(u))(t_{n,l})\|_V + \|\Pi_h\|_{\mathcal{L}(V;V_h)}\|(u - \Pi_\tau^k(u))(t_{n,l})\|_V\\ &\le \|u - \Pi_h(u)\|_{L^\infty(J;V)} + c\,\tau^{k+1}|u|_{W^{k+1,\infty}(J;V)} =: C(u).\end{aligned}$$

Invoking the boundedness of a, the Cauchy–Schwarz inequality, and since the quadrature is of order $2k$ and $\sum_{n\in\mathcal{N}_\tau}\sum_{l\in\{1:k+1\}}\omega_{n,l} = T$, we infer that

$$|\mathfrak{T}_4| \le \sum_{n\in\mathcal{N}_\tau}\sum_{l\in\{1:k+1\}}\omega_{n,l}MC(u)\|y_{h\tau}(t_{n,l})\|_V \le MT^{\frac12}C(u)\|y_{h\tau}\|_{L^2(J;V)}.$$

Combining the above estimates and recalling the definition of the $\|\cdot\|_{X_{h\tau}}$-norm and the definition of $c_1(u)$ in the assertion shows that

$$\begin{aligned}\sup_{y_{h\tau}\in X_{h\tau}}\frac{|b_\tau(e_{h\tau},y_{h\tau})|}{\|y_{h\tau}\|_{X_{h\tau}}} &\le \sqrt{2\alpha}\|u_0 - \Pi_h(u_0)\|_L + c\,\alpha\tau^{k+1}c_1(u)\\ &\quad + \|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + MT^{\frac12}\|u - \Pi_h(u)\|_{L^\infty(J;V)}.\end{aligned}$$

We now invoke the coercivity property (69.25) which implies that

$$\alpha\|e_{h\tau}\|_{X_{h\tau}} \le \sup_{y_{h\tau}\in X_{h\tau}}\frac{|b_\tau(e_{h\tau},y_{h\tau})|}{\|y_{h\tau}\|_{X_{h\tau}}}.$$

Combining the above two bounds and using the definition of ξ_κ yields

$$\begin{aligned}\|e_{h\tau}\|_{X_{h\tau}} &\le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0 - \Pi_h(u_0)\|_L + c\,\tau^{k+1}c_1(u)\\ &\quad + \frac{1}{\alpha}\|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + \xi_\kappa T^{\frac12}\|u - \Pi_h(u)\|_{L^\infty(J;V)}.\end{aligned}$$

Finally, the triangle inequality implies that $\|u - u_h\|_{X_{h\tau}} \le \|e_{h\tau}\|_{X_{h\tau}} + \|\eta\|_{X_{h\tau}}$. Using the definition of the time scales ρ and $\rho_\sharp$ yields

$$\|\eta\|_{X_{h\tau}} \le \|\eta\|_{L^2(J;V)} + c\,\frac{\iota_{L,V}}{\sqrt{\alpha}}\|\eta\|_{L^\infty(J;V)} \le c'\rho_\sharp^{\frac12}\|\eta\|_{L^\infty(J;V)}.$$

Reasoning as above then shows that $\|\eta\|_{L^\infty(J;V)} \le \|u - \Pi_h(u)\|_{L^\infty(J;V)} + c\,\tau^{k+1}|u|_{W^{k+1,\infty}(J;V)}$. Putting everything together concludes the proof. □

Remark 69.19 (Optimality in time). The identity (69.30) satisfied by the operator Π_τ^k is the key to achieve an optimal error estimate in time. □

Remark 69.20 (Supercloseness). Assume that a is time-independent so that one can use the elliptic projector $\Pi_h := \Pi_h^{\mathrm{E}}$ in the proof of Theorem 69.18. Since the operators Π_τ^k and Π_h^{E} commute and $\int_J a(\Pi_\tau^k(u) - \Pi_h^{\mathrm{E}}(\Pi_\tau^k(u)), y_{h\tau})\mathrm{d}t = 0$, we have $\int_J a(u-\eta, y_{h\tau})\mathrm{d}t = \int_J a(u - \Pi_\tau^k(u), y_{h\tau})\mathrm{d}t$. This in turn implies that $\|\int_J a(u-\eta,\cdot)\mathrm{d}t\|_{X_{h\tau}'} \le M\|u - \Pi_\tau^k(u)\|_{L^2(J;V)} \le cM\tau^{k+1}|u|_{H^{k+2}(J;V)}$. One finally obtains

$$\|e_{h\tau}\|_{X_{h\tau}} \le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0 - \Pi_h^{\mathrm{E}}(u_0)\|_L + c\,\tau^{k+1}c_1(u) + \frac{1}{\alpha}\|\partial_t u - \Pi_h^{\mathrm{E}}(\partial_t u)\|_{L^2(J;V')}.$$

This estimate delivers optimal order in space since $\|\cdot\|_{V'} \le \frac{\sqrt{\alpha\rho}}{\sqrt{2}}\|\cdot\|_L$. □

Remark 69.21 (Convergence, heat equation). Let us consider the approximation of the heat equation with H^1-conforming finite elements. Let $r \in [1, k']$, where $k' \ge 1$ is the degree of the finite elements used to build V_h. Assume that $u \in H^{k+2}(J; H^{-1}(D)) \cap W^{k+1,\infty}(J; H_0^1(D)) \cap W^{1,\infty}(J; H^{k'+1}(D))$. Then the estimate from Theorem 69.18 implies that $\|u - u_{h\tau}\|_{L^2(J;H_0^1)}$ decays as $\mathcal{O}(\tau^{k+1}c_1(u) + h^{k'}c_2(u))$ with $c_2(u) := \rho_\sharp^{\frac{1}{2}}\|u\|_{W^{1,\infty}(J;H^{k'+1})}$. Moreover, the estimate from Remark 69.20 implies that $\|(u-u_{h\tau})(T)\|_{L^2(D)}$ decays as $\mathcal{O}(\tau^{k+1}c_1(u) + h^{k'+s}\ell_D^{-s}c_2(u))$, where $s \in (0,1]$ is the elliptic regularity pickup index ($s = 1$ if there is full elliptic regularity pickup). Finally, since the constant c in the estimate does not depend on T, the error $\sup_{n\in\mathcal{N}_\tau}\|(u-u_{h\tau})(t_n)\|_{L^2(D)}$ decays with the same rate. □

Remark 69.22 (Literature). Further developments on the error analysis can be found in Thomée [273, Chap. 12]. In particular, [273, Thm. 12.2] shows that $\|u-u_{h\tau}\|_{L^\infty(\overline{J};L)} \le c\big(\tau^{k+1}(\rho|u|_{H^{k+2}(J;L)} + \sqrt{\alpha\rho}|u|_{W^{k+1,\infty}(J;V)}) + \|u - \Pi_h^{\mathrm{E}}(u)\|_{L^2(J;L)}\big)$, and under more restrictive smoothness assumptions, [273, Thm. 12.3] shows that the error in time decays as $\mathcal{O}(\tau^{2k+1})$ for $k \ge 1$. □

69.4 Algebraic realization

Let us set $m := k+1$. Recall that the quadrature induced by the mapping $T_n : \widehat{J} := [-1,1] \to \overline{J}_n$ defined in (69.1) has nodes $\{t_{n,l} := T_n(\xi_l)\}_{l\in\{1:m\}}$ and weights $\{\omega_{n,l} := \frac{\tau}{2}\omega_l\}_{l\in\{1:m\}}$. Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h, e.g., the global shape functions in the finite element space V_h (recall that these functions are also defined by invoking a mapping to a reference element, see Proposition 9.2 and §19.1). Let the mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$, the time-dependent stiffness matrices $\mathcal{A}^{n,p} \in \mathbb{R}^{I\times I}$, and the load vectors $\mathsf{F}^{n,p} \in \mathbb{R}^I$ be such that for all $p \in \{1{:}m\}$, all $n \in \mathcal{N}_\tau$, and all $i,j \in \{1{:}I\}$,

$$\mathcal{M}_{ij} := (\varphi_j, \varphi_i)_L, \quad \mathcal{A}_{ij}^{n,p} := a(t_{n,p}; \varphi_j, \varphi_i), \quad \mathsf{F}_i^{n,p} := \langle f(t_{n,p}), \varphi_i\rangle_{V',V}. \tag{69.31}$$

69.4.1 IRK implementation

Since the solution produced by the dG(k) scheme and the Radau IIA IRK scheme are identical according to Lemma 69.11, one way to implement the method is to use the IRK strategy (69.22) with $s := k+1 = m$ stages. Recall that $u_h^0 := \mathcal{P}_{V_h}(u_0)$, $u_{h\tau|J_n} = \sum_{p\in\{1:m\}} u_h^{n,p}\mathcal{L}_p \circ T_n^{-1}$, and $u_h^{n,m} = u_{h\tau}(t_n)$ for all $n \in \mathcal{N}_\tau$. We define U^n to be the coordinate vector of u_h^n in the basis $\{\varphi_i\}_{i\in\{1:I\}}$ for all $n \in \overline{\mathcal{N}}_\tau$. Likewise we define $\mathsf{U}^{n,p}$ to be the coordinate vector of $u_h^{n,p}$ in the basis

$\{\varphi_i\}_{i\in\{1:I\}}$ for all $p\in\{1{:}m\}$. Then at each time step $n\in\mathcal{N}_\tau$, (69.22) amounts to solving the following linear system:

$$\begin{pmatrix} \mathcal{M}+\tau a_{11}\mathcal{A}^{n,1} & \cdots & \tau a_{1m}\mathcal{A}^{n,m} \\ \vdots & \ddots & \vdots \\ \tau a_{m1}\mathcal{A}^{n,1} & \cdots & \mathcal{M}+\tau a_{mm}\mathcal{A}^{n,m} \end{pmatrix} \begin{pmatrix} \mathsf{U}^{n,1} \\ \vdots \\ \mathsf{U}^{n,m} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathsf{G}}^{n,1} \\ \vdots \\ \widetilde{\mathsf{G}}^{n,m} \end{pmatrix}, \tag{69.32}$$

with the coefficients $\{a_{pq}\}_{p,q\in\{1:m\}}$ defined in (69.23) and the load vectors $\widetilde{\mathsf{G}}^{n,p}:=\mathcal{M}\mathsf{U}^{n-1}+\tau\sum_{q\in\{1:m\}}a_{pq}\mathsf{F}^{n,q}\in\mathbb{R}^I$. Finally, we set $\mathsf{U}^n:=\mathsf{U}^{n,m}$.

69.4.2 General case

We now write the linear system corresponding to the dG(k) scheme (69.16) for a general basis $\{\psi_q\}_{q\in\{1:m\}}$ of $\mathbb{P}_k(\widehat{J};\mathbb{R})$. For all $n\in\mathcal{N}_\tau$ and all $q\in\{1{:}m\}$, we introduce the coordinate vectors $\mathsf{U}^{n,q}\in\mathbb{R}^I$ s.t. $u_{h\tau}(\boldsymbol{x},t):=\sum_{j\in\{1:I\}}\sum_{q\in\{1:m\}}\mathsf{U}_j^{n,q}\psi_q(T_n^{-1}(t))\varphi_j(\boldsymbol{x})$ for all $(\boldsymbol{x},t)\in D\times J_n$. For all $n\in\mathcal{N}_\tau$ an all $p,q\in\{1{:}m\}$, we define

$$\mathcal{A}^{n,pq}:=\sum_{l\in\{1:m\}}\frac{\omega_l}{2}\psi_q(\xi_l)\psi_p(\xi_l)\mathcal{A}^{n,l}\in\mathbb{R}^{I\times I}, \tag{69.33}$$

$$b_{pq}:=\int_{-1}^1\psi_q'(\xi)\psi_p(\xi)\,\mathrm{d}\xi+\psi_q(-1)\psi_p(-1). \tag{69.34}$$

Considering test functions of the form $\varphi_i(\boldsymbol{x})\psi_p(T_n^{-1}(t))$, we rewrite the dG($k$) scheme (69.16) in the following block form: For all $n\in\mathcal{N}_\tau$,

$$\begin{pmatrix} b_{11}\mathcal{M}+\tau\mathcal{A}^{n,11} & \cdots & b_{1m}\mathcal{M}+\tau\mathcal{A}^{n,1m} \\ \vdots & \ddots & \vdots \\ b_{m1}\mathcal{M}+\tau\mathcal{A}^{n,m1} & \cdots & b_{mm}\mathcal{M}+\tau\mathcal{A}^{n,mm} \end{pmatrix} \begin{pmatrix} \mathsf{U}^{n,1} \\ \vdots \\ \mathsf{U}^{n,m} \end{pmatrix} = \begin{pmatrix} \mathsf{G}^{n,1} \\ \vdots \\ \mathsf{G}^{n,m} \end{pmatrix}, \tag{69.35}$$

and set $\mathsf{U}^n:=\sum_{p\in\{1:m\}}\psi_p(1)\mathsf{U}^{n,p}$. The load vectors are defined by $\mathsf{G}^{n,p}:=\psi_p(-1)\mathcal{M}\mathsf{U}^{n-1}+\tau\sum_{q\in\{1:m\}}\frac{\omega_q}{2}\psi_p(\xi_q)\mathsf{F}^{n,q}\in\mathbb{R}^I$ for all $p\in\{1{:}m\}$.

The algorithmic complexity of the time-stepping schemes (69.32) and (69.35) is a priori high since one has to assemble at each time step $m=k+1$ stiffness matrices, and one has to solve a globally coupled linear system of size $I\times m$. Moreover, even if the bilinear form a is symmetric, the system matrices in (69.32) and (69.35) are nonsymmetric for all $k\geq 1$. If the bilinear form a is time-independent, the assembling of the matrix on the left-hand side of (69.35) simplifies since $\mathcal{A}^{n,pq}=m_{pq}\mathcal{A}$ with the time-independent stiffness matrix $\mathcal{A}\in\mathbb{R}^{I\times I}$ and the coefficients m_{pq} given by

$$\mathcal{A}_{ij}:=a(\varphi_j,\varphi_i),\qquad m_{pq}:=\frac{1}{2}\int_{-1}^1\psi_q(s)\psi_p(s)\,\mathrm{d}s, \tag{69.36}$$

for every $i,j\in\{1{:}I\}$ and $p,q\in\{1{:}m\}$, that is, it is only necessary to assemble *one* stiffness matrix. The same remark holds for (69.32).

The algebraic formulations (69.32) and (69.35) can be rewritten in a more compact form using tensor notation. Let us focus on (69.35). We introduce the matrices $\mathbb{B},\mathbb{M}\in\mathbb{R}^{m\times m}$ with entries $\mathbb{B}_{pq}:=b_{pq}$ and $\mathbb{M}_{pq}:=m_{pq}$. Consider the notation $(\mathcal{M}\otimes\mathbb{B})_{ip,jq}:=\mathcal{M}_{ij}b_{pq}$ for all $i,j\in\{1{:}I\}$ and $p,q\in\{1{:}m\}$, where $ip\in\{1{:}m{\times}I\}$ abbreviates $(p-1)I+i$ and $jq\in\{1{:}m{\times}I\}$ abbreviates $(q-1)I+j$. Using the same notation for $\mathcal{A}\otimes\mathbb{M}$, we can rewrite (69.35) as follows:

$$\boldsymbol{\mathcal{S}}\mathbf{U}^n=\mathbf{G}^n,\qquad\boldsymbol{\mathcal{S}}:=\mathcal{M}\otimes\mathbb{B}+\tau\mathcal{A}\otimes\mathbb{M}\in\mathbb{R}^{Im\times Im}, \tag{69.37}$$

with $\mathsf{U}^n := (\mathsf{U}^{n,1},\dots,\mathsf{U}^{n,m})^\mathsf{T} \in \mathbb{R}^{Im}$ and $\mathsf{G}^n := (\mathsf{G}^{n,1},\dots,\mathsf{G}^{n,m})^\mathsf{T} \in \mathbb{R}^{Im}$. Notice that the matrix $\mathbb{M}$ is diagonal if one chooses $\{\psi_l := \mathcal{L}_l\}_{l\in\{1:m\}}$.

Remark 69.23 (Symmetrization, preconditioning). Assume that the bilinear form a is time-independent so that the linear system at each time step takes the form (69.37). Assume that a is symmetric so that the stiffness matrix $\mathcal{A}$ is symmetric (recall that the mass matrix is by construction symmetric). Observe that the system matrix $\boldsymbol{\mathcal{S}}$ in (69.37) is nonsymmetric owing to the lack of symmetry of the matrix $\mathbb{B}$ (recall that $\mathbb{M}$ is by construction symmetric). An interesting option to symmetrize (69.37) is to precondition it on the left by the matrix $\boldsymbol{\mathcal{P}} := \boldsymbol{\mathcal{S}}^\mathsf{T}(\mathcal{A}^{-1}\otimes\frac{1}{\tau}\mathbb{M}^{-1})$ leading to the preconditioned symmetric linear system $\hat{\boldsymbol{\mathcal{S}}}\mathsf{U}^n = \mathsf{H}^n$ with

$$\hat{\boldsymbol{\mathcal{S}}} := \boldsymbol{\mathcal{S}}^\mathsf{T}(\mathcal{A}^{-1}\otimes\tfrac{1}{\tau}\mathbb{M}^{-1})\boldsymbol{\mathcal{S}}, \qquad \mathsf{H}^n := \boldsymbol{\mathcal{S}}^\mathsf{T}(\mathcal{A}^{-1}\otimes\tfrac{1}{\tau}\mathbb{M}^{-1})\mathsf{G}^n. \tag{69.38}$$

Recalling that $(\mathcal{C}\otimes\mathbb{X})^\mathsf{T} = \mathcal{C}^\mathsf{T}\otimes\mathbb{X}^\mathsf{T}$ and $(\mathcal{C}\otimes\mathbb{X})(\mathcal{D}\otimes\mathbb{Y}) = (\mathcal{C}\mathcal{D}\otimes\mathbb{X}\mathbb{Y})$, a straightforward calculation shows that

$$\hat{\boldsymbol{\mathcal{S}}} = \tfrac{1}{\tau}(\mathcal{M}\mathcal{A}^{-1}\mathcal{M})\otimes(\mathbb{B}^\mathsf{T}\mathbb{M}^{-1}\mathbb{B}) + \mathcal{M}\otimes(\mathbb{B}+\mathbb{B}^\mathsf{T}) + \tau\mathcal{A}\otimes\mathbb{M}.$$

Let $\mathcal{R}_n : \mathbb{P}_k(J_n;H)\to\mathbb{P}_{k+1}(J_n;H)$ be s.t. $\mathcal{R}_n(v) := v - v(t_{n-1}^+)\theta_{k+1}\circ T_n^{-1}$ for all $v\in\mathbb{P}_k(J_n;H)$ and all $n\in\mathcal{N}_\tau$, with θ_{k+1} defined in Remark 69.9. One can show that $\hat{\boldsymbol{\mathcal{S}}}$ is the stiffness matrix of the least-squares minimization of the residual norm $\|A_h^{-1}(\partial_t\mathcal{R}_n(v_{h\tau})) + v_{h\tau}\|^2_{L^2(J_n;V_h)}$, with the inner product $(v_{hn},w_{hn})_{L^2(J_n;V_h)} := \int_{J_n}(A_h(v_{hn}),w_{hn})_L\mathrm{d}t$ for all $v_{hn},w_{hn}\in\mathbb{P}_k(J_n;V_h)$. The least-squares minimization viewpoint is adopted, e.g., in Nouy [231], Andreev [9, 10], Boiveau et al. [37]. We refer the reader to Smears [262] for further insight on how to precondition efficiently the symmetric system (69.38). See also Exercise 69.8. □

Exercises

Exercise 69.1 (Integral identities). Prove the identities (69.11). (*Hint*: use that the Gauss–Radau quadrature is of order $2k$.)

Exercise 69.2 (Equivalence with Radau IIA IRK). Prove the converse assertion in Lemma 69.11. (*Hint*: show that

$$\mathcal{R}_\tau(u_{h\tau})(t) = u_h^{n-1} + \tau\sum_{j\in\{1:k+1\}}\frac{1}{2}\int_{-1}^{T_n^{-1}(t)}\mathcal{L}_j(\xi)\,\mathrm{d}\xi\big(f_h(t_{n,j}) - \mathcal{A}_h(t_{n,j})(u_h^{n,j})\big),$$

for all $t\in J_n$.)

Exercise 69.3 (Poincaré in time). Let $n\in\mathcal{N}_\tau$ and H be a Hilbert space. Show that $\|v\|^2_{L^2(J_n;H)} \le 2\tau\|v(t_{n-1}^+)\|_H^2 + \tau^2\|\partial_t v\|^2_{L^2(J_n;H)}$ for all $v\in H^1(J_n;H)$. (*Hint*: use that $v(t) = v(t_{n-1}^+) + \int_{t_{n-1}}^t\partial_t v\mathrm{d}t$ for all $t\in J_n$.)

Exercise 69.4 (Time reconstruction). (i) Show that the definition of R_τ given in Remark 69.9 is equivalent to Definition 69.5. (ii) Show that the two definitions of θ_{k+1} given in Remark 69.9 are identical. (*Hint*: set $\delta(s) := \frac{(-1)^k}{2}(L_k - L_{k+1}) - \prod_{l\in\{1:k+1\}}\frac{\xi_l - s}{\xi_l+1}$ and prove that $\delta(-1) = 0$ and $\int_{\widehat{J}}\delta'(s)q(s)\,\mathrm{d}s = 0$ for all $q\in\mathbb{P}_k(\widehat{J};\mathbb{R})$.) (iii) Let $(V, L\equiv L', V')$ be a Gelfand triple. Let $\widehat{\mathcal{R}} : \mathbb{P}_k(\widehat{J};\mathbb{R})\to\mathbb{P}_{k+1}(\widehat{J};\mathbb{R})$ be s.t. $\widehat{\mathcal{R}}(q) := q - q(-1)\theta_{k+1}$. Let $\mathcal{R}_n : \mathbb{P}_k(J_n;\mathbb{R})\to\mathbb{P}_{k+1}(J_n;\mathbb{R})$ be

s.t. $\mathcal{R}_n(v) = \sum_{q\in\{1:k+1\}} \mathsf{V}_q \widehat{\mathcal{R}}(\psi_q)\circ T_n^{-1}$ for all $v := \sum_{q\in\{1:k+1\}} \mathsf{V}_q\psi_q\circ T_n^{-1}$ and all $n\in\mathcal{N}_\tau$, where $\{\psi_q\}_{q\in\{1:k+1\}}$ is a basis for $\mathbb{P}_k(\widehat{J};\mathbb{R})$. Accept as a fact that $\|v\|_{L^\infty(J_n;V')} \le 2^{2-\frac{1}{p}}\|\partial_t\mathcal{R}_n(v)\|_{L^p(J_n;V')}$ for all $p\in[1,\infty]$ and all $v\in\mathbb{P}_k(J_n;V')$ (see Holm and Wihler [185, Prop. 1]). Prove that $\|v\|_{L^2(J_n;L)} \le (2\tau)^{\frac12}\|\partial_t\mathcal{R}_n(v)\|^{\frac12}_{L^2(J_n;V')}\|v\|^{\frac12}_{L^2(J_n;V)}$ for all $v\in\mathbb{P}_k(J_n;V)$ and all $n\in\mathcal{N}_\tau$. (*Hint*: $\|\phi\|_L^2 \le \|\phi\|_{V'}\|\phi\|_V$ for all $\phi\in V$.)

Exercise 69.5 (dG(1)). Assume that a is time-independent. (i) Verify that the dG(1) scheme amounts to

$$\begin{pmatrix} \frac{9}{8}\mathcal{M} & \frac{3}{8}\mathcal{M} \\ -\frac{9}{8}\mathcal{M} & \frac{5}{8}\mathcal{M}\end{pmatrix}\begin{pmatrix}\mathsf{U}^{n,1}\\ \mathsf{U}^{n,2}\end{pmatrix} + \tau\begin{pmatrix}\frac{3}{4}\mathcal{A}\mathsf{U}^{n,1}\\ \frac14\mathcal{A}\mathsf{U}^{n,2}\end{pmatrix} = \begin{pmatrix}\frac32\mathcal{M}\mathsf{U}^{n-1}\\ -\frac12\mathcal{M}\mathsf{U}^{n-1}\end{pmatrix} + \tau\begin{pmatrix}\frac34\mathsf{F}^{n,1}\\ \frac14\mathsf{F}^{n,2}\end{pmatrix},$$

and $\mathsf{U}^n = \mathsf{U}^{n,2}$, where $\mathsf{U}^{n,1}$ and $\mathsf{U}^{n,2}$ are the coordinate vectors of the discrete solution at $t_{n-1}+\frac13\tau$ and at t_n, respectively. (*Hint*: use the Lagrange interpolation polynomials associated with the two Gauss–Radau nodes $\xi_1 := -\frac13$ and $\xi_2 := 1$.) (ii) Using the same notation as above, write the scheme in IRK form. (*Hint*: see (69.22) and (69.24).)

Exercise 69.6 (IRK final stage). The objective of this exercise is to prove the assertions made in Remark 69.13. (i) Show that for every s-stage IRK scheme, the update u_h^n is given by $u_h^n = \alpha_0 u_h^{n-1} + \sum_{p\in\{1:s\}}\alpha_p u_h^{n,p}$, where $\alpha_p := \sum_{q\in\{1:s\}} b_q(a^{-1})_{qp}$, $\alpha_0 := 1-\sum_{p\in\{1:s\}}\alpha_p$, and $(a^{-1})_{pq}$ are the coefficients of the inverse of the Butcher matrix $(a_{pq})_{p,q\in\{1:s\}}$. (ii) Show that for the Radau IIA IRK scheme, $\alpha_p = 0$ for all $p\in\{0:s-1\}$ and $\alpha_s = 1$.

Exercise 69.7 (Π_τ^k). (i) Prove the uniform stability of Π_n^k in $L^\infty(J_n;Z)$ with $Z\subseteq L$. (*Hint*: map to the reference interval $\widehat{J}$.) Prove (69.27). (*Hint*: accept as a fact that the standard polynomial approximation properties in Sobolev spaces extend to Bochner spaces.) (ii) Build the operator Π_n^k with $Z := V'$ as in Remark 69.17. (*Hint*: use the Riesz–Fréchet operator $J^{\mathrm{RF}} : L^2(J_n;V)\to (L^2(J_n;V))' = L^2(J_n;V')$.) Adapt the identity in Lemma 69.16 to the case $Z := V'$. (*Hint*: invoke the integration by parts formula (64.7).) Prove a stability estimate for Π_n^k in $L^\infty(J_n;V')$. (iii) Let $\Pi_h\in\mathcal{L}(V;V_h)$. Show that $\delta := \Pi_\tau^k(\Pi_h(v)) - \Pi_h(\Pi_\tau^k(v)) = 0$ for all $v\in H^1(J;V)$. (*Hint*: show that $\delta(t_n)=0$ for all $n\in\overline{\mathcal{N}}_\tau$ and that $\int_{J_n}(\delta,q)_L\mathrm{d}t = 0$ for all $q\in\mathbb{P}_{k-1}(J_n;V_h)$ and all $n\in\mathcal{N}_\tau$.)

Exercise 69.8 (Symmetrization). Let $\widehat{\mathcal{R}}$ be defined in Exercise 69.4(iii). (i) Prove that $\mathbb{B}_{pq} = \int_{-1}^1 \widehat{\mathcal{R}}(\psi_q)'\psi_p\,\mathrm{d}s$, $(\mathbb{B}+\mathbb{B}^{\mathsf{T}})_{pq} = \psi_q(-1)\psi_p(-1)+\psi_q(1)\psi_p(1)$, $(\mathbb{B}^{\mathsf{T}}\mathbb{M}^{-1}\mathbb{B})_{pq} = \int_{-1}^1\widehat{\mathcal{R}}(\psi_q)'\widehat{\mathcal{R}}(\psi_p)'\,\mathrm{d}s$ for all $p,q\in\{1:m\}$. (*Hint*: use Exercise 28.1.) (ii) Set $\hat{\boldsymbol{\mathcal{S}}}_\flat := \frac1\tau(\mathcal{M}\mathcal{A}^{-1}\mathcal{M})\otimes(\mathbb{B}^{\mathsf{T}}\mathbb{M}^{-1}\mathbb{B}) + \tau\mathcal{A}\otimes\mathbb{M}$. Prove that $\mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V} \le \mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}\mathsf{V} \le 2\mathsf{V}^{\mathsf{T}}\hat{\boldsymbol{\mathcal{S}}}_\flat\mathsf{V}$ for all $\mathsf{V}\in\mathbb{R}^{Im}$. (*Hint*: note that $\mathsf{V}^{\mathsf{T}}(\mathcal{M}\otimes\mathbb{B})\mathsf{V} = \mathsf{Y}^{\mathsf{T}}(\mathcal{A}^{-1}\otimes\mathbb{M}^{-1})\mathsf{Z}$ with $\mathsf{Y} := (\mathcal{A}\otimes\mathbb{M})\mathsf{V}$ and $\mathsf{Z} := (\mathcal{M}\otimes\mathbb{B})\mathsf{V}$ and apply the Cauchy–Schwarz and Young's inequalities.) (iii) Verify that $\hat{\boldsymbol{\mathcal{S}}}$ is the stiffness matrix associated with the minimization of the residual norm $\|A_h^{-1}(\partial_t\mathcal{R}_n(v_{h\tau}))+v_{h\tau}\|^2_{L^2(J_n;V_h)}$. (*Hint*: use again Exercise 28.1.) (iv) Compute the matrix $\hat{\boldsymbol{\mathcal{S}}}$ for $k := 1$. (*Hint*: see Exercise 69.5.)

Chapter 70

Continuous Petrov–Galerkin in time

In this chapter, we continue the study started in the previous chapter on higher-order time approximation schemes using a space-time functional framework. Recall that the test functions are discontinuous at the time nodes so as to obtain a time-stepping procedure. In the previous chapter, the trial functions are also discontinuous at the time nodes, and the resulting method is a discontinuous Galerkin scheme in time. In the present chapter, the trial functions are continuous in time and piecewise polynomials with a polynomial degree that is one order higher than that of the test functions. The resulting technique is called continuous Petrov–Galerkin method, and its lowest-order version is the Crank–Nicolson scheme studied in §68.3. Like the dG(k) schemes, the continuous Petrov–Galerkin schemes are implicit one-step methods. They can also be interpreted as implicit Runge–Kutta methods.

70.1 Formulation of the method

We describe the continuous Petrov–Galerkin method in this section. We use the notation as in §69.1 for the time discretization.

70.1.1 Quadratures and interpolation

Let $k \geq 1$ be the polynomial degree for the time discretization. Let $\{\xi_l\}_{l\in\{1:k\}}$ be the *Gauss–Legendre nodes* in the reference interval $\widehat{J} := (-1, 1]$, and let $\{\omega_l\}_{l\in\{1:k\}}$ be the corresponding weights. This set of nodes and weights gives a quadrature of order $(2k-1)$ (see §6.2). Notice that in the previous chapter on dG(k) schemes, we used $k \geq 0$ and $(k+1)$ Gauss–Radau nodes for the quadrature. Using the mapping $T_n : \widehat{J} \to J_n$ defined in (69.1), we obtain a quadrature in J_n with $t_{n,l} := T_n(\xi_l)$ and $\omega_{n,l} := \frac{\tau}{2}\omega_l$ for all $l \in \{1:k\}$. We introduce the discrete measure $\mu_k^{\mathrm{GL}}(\mathrm{d}t)$ defined on J by setting

$$\int_J g(t)\,\mu_k^{\mathrm{GL}}(\mathrm{d}t) := \sum_{n\in\mathcal{N}_\tau} \int_{J_n} g(t)\,\mu_k^{\mathrm{GL}}(\mathrm{d}t) := \sum_{n\in\mathcal{N}_\tau} \sum_{l\in\{1:k\}} \omega_{n,l}\, g(t_{n,l}), \tag{70.1}$$

for all $g \in C^0(\overline{J};\mathbb{R})$. Notice that we slightly abuse the terminology by using the same symbol for the discrete measure on J and its restriction on the time interval J_n.

Let $Z \in \{V', L, V_h\}$ and recall that $H^1(J;Z) \hookrightarrow C^0(\overline{J};Z)$. Let $\mathcal{I}_{k-1}^{\mathrm{GL}} : H^1(J;Z) \to P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;Z)$ be the Lagrange *interpolation operator* associated with the Gauss–Legendre nodes such that for all $v \in H^1(J;Z)$, $\mathcal{I}_{k-1}^{\mathrm{GL}}(v)(0) := v(0)$ and for all $n \in \mathcal{N}_\tau$,

$$\mathcal{I}_{k-1}^{\mathrm{GL}}(v)_{|J_n} := \sum_{l\in\{1:k\}} v(t_{n,l})\mathcal{L}_l \circ T_n^{-1}, \tag{70.2}$$

where $\mathcal{L}_l(\xi) := \prod_{j\in\{1:k\}\setminus\{l\}} \frac{\xi-\xi_j}{\xi_l-\xi_j} \in \mathbb{P}_{k-1}(\widehat{J};\mathbb{R})$, i.e., $\mathcal{L}_l(\xi_{l'}) = \delta_{ll'}$ for all $l, l' \in \{1:k\}$. In view of the error analysis, we observe that $\mathcal{I}_{k-1}^{\mathrm{GL}}$ does not have optimal approximation properties since it is a piecewise polynomial of degree $(k-1)$ in time. This motivates the introduction of another Lagrange interpolation operator based on the Gauss–Legendre nodes and one of the two endpoints of each time interval (we choose the right one to fix the ideas). Recalling that L_k is the k-th Legendre polynomial, we define $\mathcal{I}_k^{\mathrm{GL}+} : H^1(J;Z) \to P_k^{\mathrm{b}}(\overline{J}_\tau;Z)$ by setting for all $v \in H^1(J;Z)$, $\mathcal{I}_k^{\mathrm{GL}+}(v)(0) = v(0)$ and for all $n \in \mathcal{N}_\tau$,

$$\mathcal{I}_k^{\mathrm{GL}+}(v)_{|J_n} = v(t_n)L_k \circ T_n^{-1} + \sum_{l\in\{1:k\}} u(t_{n,l})\frac{t-t_n}{t_{n,l}-t_n}\mathcal{L}_l \circ T_n^{-1}. \tag{70.3}$$

Since $\mathcal{I}_k^{\mathrm{GL}+}$ is $L^\infty(J;Z)$-stable uniformly w.r.t. τ and leaves $P_k^{\mathrm{b}}(J_\tau;Z)$ pointwise invariant, there is c such that for all τ and all $v \in H^{k+1}(J;Z)$,

$$\|v - \mathcal{I}_k^{\mathrm{GL}+}(v)\|_{L^2(J;Z)} \le c\,\tau^{k+1}|v|_{H^{k+1}(J;Z)}. \tag{70.4}$$

Moreover, the following identity holds true for all $v \in H^1(J;L)$ and all $y_\tau \in P_{k-1}^{\mathrm{b}}(J_\tau;L)$ (see Exercise 70.1):

$$\int_J (v, y_\tau)_L \mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_J (\mathcal{I}_k^{\mathrm{GL}+}(v), y_\tau)_L \mathrm{d}t. \tag{70.5}$$

70.1.2 Discretization in time

The idea behind the *continuous Petrov–Galerkin* cPG(k) time scheme is to consider a trial space composed of continuous, piecewise polynomial functions in time of degree k and a test space composed of discontinuous, piecewise polynomial functions of degree $(k-1)$. This leads to a conforming approximation in time. (Recall that the approximation setting is nonconforming for the dG(k) schemes studied in Chapter 69.)

The time-discrete trial and test spaces are taken to be

$$X_{h\tau} := P_k^{\mathrm{g}}(\overline{J}_\tau;V_h), \qquad Y_{h\tau} := P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;V_h). \tag{70.6}$$

We consider the bilinear form b_τ such that for all $(v_{h\tau}, y_{h\tau}) \in X_{h\tau}\times Y_{h\tau}$,

$$\begin{aligned} b_\tau(v_{h\tau}, y_{h\tau}) := {} & (v_{h\tau}(0), y_{h\tau}(0))_L \\ & + \int_J (\partial_t v_{h\tau}(t), y_{h\tau}(t))_L \mathrm{d}t + \int_J a(t; v_{h\tau}(t), y_{h\tau}(t))\,\mu_k^{\mathrm{GL}}(\mathrm{d}t). \end{aligned} \tag{70.7}$$

Notice that the time derivative of $v_{h\tau}$ is integrable over J since $v_{h\tau}$ is continuous in time by construction. Observe also that if the bilinear form a is time-independent, we have for all $n \in \mathcal{N}_\tau$,

$$\int_{J_n} a(v_{h\tau}(t), y_{h\tau}(t))\,\mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_{J_n} a(v_{h\tau}(t), y_{h\tau}(t))\mathrm{d}t,$$

since the integrand is in $\mathbb{P}_{2k-1}(J_n;\mathbb{R})$ and the quadrature is of order $(2k-1)$. Similarly, we consider the linear form ℓ_τ such that for all $y_{h\tau} \in Y_{h\tau}$,

$$\ell_\tau(y_{h\tau}) := (u_0, y_{h\tau}(0))_L + \int_J \langle f(t), y_{h\tau}(t)\rangle_{V',V}\, \mu_k^{\mathrm{GL}}(\mathrm{d}t).$$

The cPG(k) scheme consists of the following space-time discrete problem:

$$\begin{cases} \text{Find } u_{h\tau} \in X_{h\tau} \text{ such that} \\ b_\tau(u_{h\tau}, y_{h\tau}) = \ell_\tau(y_{h\tau}), \quad \forall y_{h\tau} \in Y_{h\tau}. \end{cases} \tag{70.8}$$

This problem amounts to solving a square linear system since $\dim(X_{h\tau}) = \dim(Y_{h\tau}) = (1+Nk)\times \dim(V_h)$. The size of this linear system is smaller (for fixed k) than that induced by the dG(k) scheme. As we shall see below, the price to pay for this slight reduction in complexity is that the cPG(k) schemes have somewhat weaker stability properties than the dG(k) schemes.

Proposition 70.1 (Localization). *The* cPG(k) *solution $u_{h\tau}$ (if it exists) is s.t. $u_{h\tau}(0) = \mathcal{P}_{V_h}(u_0)$ and for all $q \in \mathbb{P}_{k-1}(J_n;V_h)$ and all $n \in \mathcal{N}_\tau$,*

$$\int_{J_n} (\partial_t u_{h\tau}(t), q(t))_L \mathrm{d}t + \int_{J_n} a(t; u_{h\tau}(t), q(t))\, \mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_{J_n} \langle f(t), q(t)\rangle_{V',V}\, \mu_k^{\mathrm{GL}}(\mathrm{d}t). \tag{70.9}$$

Proof. Proceed as in the proof of Proposition 69.2. □

Proposition 70.1 shows that the cPG(k) scheme gives a time-stepping procedure, where one first sets $u_{h\tau}(0) := \mathcal{P}_{V_h}(u_0)$ and then one computes sequentially the restrictions $u_{h\tau|J_n}$ by solving (70.9) for $n = 1, 2, \dots, N$.

Recalling that $\mathcal{P}_{V_h} : L \to V_h$ is the L-orthogonal projection from L onto V_h, let us set $f_h(t) := \mathcal{P}_{V_h}(f(t)) \in V_h$ for all $t \in J$. We also define $A_h(t) : V_h \to V_h$ s.t. $(A_h(t)(v_h), w_h)_L := a_h(t; v_h, w_h)$ for all $v_h, w_h \in V_h$ and $t \in J$.

Proposition 70.2 (Reformulation). *The scheme* (70.9) *is equivalent to the following: $u_h(0) = \mathcal{P}_{V_h}(u_0)$ and for all $l \in \{1{:}k\}$ and all $n \in \mathcal{N}_\tau$,*

$$\partial_t u_{h\tau}(t_{n,l}) + A_h(t_{n,l})(u_h(t_{n,l})) = f_h(t_{n,l}). \tag{70.10}$$

Proof. Proceed as in the proof of Proposition 69.7. □

Example 70.3 (Crank–Nicolson, cPG(1)). Let us take $k := 1$. This means that $u_{h\tau}$ is continuous and piecewise affine in time. Let us write $u_h^n := u_{h\tau}(t_n)$ for all $n \in \overline{\mathcal{N}}_\tau$, so that $u_{h\tau}(t) = \frac{t_n - t}{\tau} u_h^{n-1} + \frac{t - t_{n-1}}{\tau} u_h^n$ for all $t \in \overline{J}_n$, and $\partial_t u_{h\tau}(t) = \frac{u_h^n - u_h^{n-1}}{\tau}$ on J_n. Since the test function q_{hn} in (70.9) is constant in time over J_n, and since the 1-point Gauss–Legendre quadrature is the midpoint rule, letting $t_{n-\frac12} := \frac12(t_{n-1}+t_n)$ we obtain $(u_h^n - u_h^{n-1}, w_h)_L + \tau a(t_{n-\frac12}; \frac12(u_h^{n-1} + u_h^n), w_h) = \tau\langle f(t_{n-\frac12}), w_h\rangle_{V',V}$ for all $w_h \in V_h$, i.e., we recover the *Crank–Nicolson scheme* studied in §68.3. □

Remark 70.4 (Literature). Continuous Petrov–Galerkin methods have been studied by Hulme [191], Aziz and Monk [19]. We also refer the reader to Wihler [286], Schötzau and Wihler [249], Hussain et al. [193], Ahmed and Matthies [2], Bause et al. [26], Holm and Wihler [185] for other results. □

70.1.3 Equivalence with Kuntzmann–Butcher IRK

We show in this section that the cPG(k) scheme (70.9) (or (70.10)) is equivalent to an implicit Runge–Kutta (IRK) scheme often called *Kuntzmann–Butcher (KB) method* in the literature (see Butcher [77, §3], [78, §5], Hairer and Wanner [175, §IV.5], [176, §II.7], Kuntzmann [208]).

Referring to §69.2.4, we consider a s-stage IRK scheme with the Butcher tableau s.t.

$$a_{ij} := \frac{1}{2}\int_{-1}^{\xi_i} \mathcal{L}_j(\xi)\,\mathrm{d}\xi, \qquad b_i := \frac{1}{2}\int_{-1}^{1} \mathcal{L}_i(\xi)\,\mathrm{d}\xi, \qquad c_i := \frac{\xi_i+1}{2}, \tag{70.11}$$

for all $i,j \in \{1\!:\!s\}$, where $\{\xi_i\}_{i\in\{1:s\}}$ are the Gauss–Legendre quadrature points and $\mathcal{L}_i(\xi) := \prod_{j\in\{1:s\}\setminus\{i\}} \frac{\xi_j-\xi}{\xi_j-\xi_i} \in \mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$ is the Lagrange polynomial based on these nodes and associated with the i-th node. This leads to the following time-stepping technique to approximate in time the semi-discrete problem (69.8): One first sets $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then for all $n \in \mathcal{N}_\tau$, one seeks $\{u_h^{n,j}\}_{j\in\{1:s\}} \subset V_h$ such that for all $i \in \{1\!:\!s\}$,

$$u_h^{n,i} - u_h^{n-1} = \tau \sum_{j\in\{1:s\}} a_{ij}\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big), \tag{70.12}$$

with $t_{n,j} := t_{n-1} + c_j\tau = T_n(\xi_j)$ for all $j \in \{1\!:\!s\}$, and finally one sets

$$u_h^n := u_h^{n-1} + \tau \sum_{j\in\{1:s\}} b_j\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big). \tag{70.13}$$

The expression (70.13) is not very convenient to compute u_h^n, and a better way mentioned in Remark 69.13 (to be justified in Lemma 70.5 below) is

$$u_h^n = \alpha_0 u_h^{n-1} + \sum_{l\in\{1:s\}} \alpha_l u_h^{n,l}, \qquad \alpha_0 := (-1)^s, \qquad \alpha_l := \frac{2\mathcal{L}_l(1)}{\xi_l+1}. \tag{70.14}$$

The Butcher tableaux of the one-stage, the two-stage, and the three-stage KB IRK schemes are as follows (see also [77, Tab. 2]):

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} \frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{10-3\sqrt{15}}{45} & \frac{25-6\sqrt{15}}{180} \\ \frac{1}{2} & \frac{10+3\sqrt{15}}{72} & \frac{2}{9} & \frac{10-3\sqrt{15}}{72} \\ \frac{5+\sqrt{15}}{10} & \frac{25+6\sqrt{15}}{180} & \frac{10+3\sqrt{15}}{45} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array} \tag{70.15}$$

The corresponding coefficients $(\alpha_l)_{l\in\{0:s\}}$ are $(-1,2)$ for $s=1$, $(1,-\sqrt{3},\sqrt{3})$ for $s=2$, and $(-1,\frac{5}{3},-\frac{4}{3},\frac{5}{3})$ for $s=3$. Notice that the one-stage scheme is nothing but the Crank–Nicolson method since we have $u_h^{n,1} = \frac{1}{2}(u_h^{n-1}+u_h^n)$ so that $u_h^n = u_h^{n-1} + \tau\big(f_h(t_{n-\frac{1}{2}}) - A_h(t_{n-\frac{1}{2}})(\frac{1}{2}(u_h^{n-1}+u_h^n))\big)$.

Lemma 70.5 (cPG(k) $\Leftrightarrow$ KB IRK). *Let $k \geq 1$. Let $u_{h\tau} \in X_{h\tau} := P_k^{\mathrm{g}}(\overline{J}_\tau;V_h)$ and set $u_h^n := u_{h\tau}(t_n)$ for all $n \in \overline{\mathcal{N}}_\tau$, and $\{u_h^{n,l} := u_{h\tau}(t_{n,l})\}_{l\in\{1:k\}}$ for all $n \in \mathcal{N}_\tau$.* (i) *$u_{h\tau}$ solves* (70.9) *iff $\{u_h^{n,l}\}_{l\in\{1:s\}}$ solves* (70.12) *with $s := k$ and u_h^n is given by* (70.13) *for all $n \in \mathcal{N}_\tau$.* (ii) *$u_h^n := u_{h\tau}(t_n)$ is also given by* (70.14) *for all $n \in \mathcal{N}_\tau$.*

Proof. We argue as in the proof of Lemma 69.11. Assume that $u_{h\tau}$ solves (70.9), or equivalently (70.10) owing to Proposition 70.2. Since $\partial_t u_{h\tau|J_n} \in \mathbb{P}_{k-1}(J_n;V_h)$, we infer that for all $n \in \mathcal{N}_\tau$,

$$\partial_t u_{h\tau} = \sum_{l\in\{1:k\}} \big(f_h(t_{n,l}) - A_h(t_{n,l})(u_h^{n,l})\big)\mathcal{L}_l \circ T_n^{-1}.$$

Integrating this identity over $(t_{n-1}, t_{n,i})$ for all $i \in \{1{:}k\}$, using the definition of a_{ij} in (70.11), and since $t_{n,j} = T_n(\xi_j)$, this gives

$$u_h^{n,i} - u_h^{n-1} = \tau \sum_{l\in\{1:k\}} a_{il}\big(f_h(t_{n,l}) - A_h(t_{n,l})(u_h^{n,l})\big),$$

since $u_h^{n-1} := u_{h\tau}(t_{n-1})$ and $u_h^{n,i} := u_{h\tau}(t_{n,i})$. This is exactly (70.12). Moreover, using $q(t) = 1$ in (70.9), we obtain

$$u_h^n = u_h^{n-1} + \tau \sum_{l\in\{1:k\}} \frac{1}{2}\omega_l\big(f_h(t_{n,l}) - A_h(t_{n,l})(u_h^{n,l})\big).$$

But recalling that $\omega_l = \int_{-1}^1 \mathcal{L}_l(t)\mathrm{d}t$ (see Lemma 6.4), we have $b_l = \frac{1}{2}\omega_l$ for all $l \in \{1{:}k\}$. The above identity is (70.13). This shows that (70.9) implies (70.12)-(70.13). The converse is established in Exercise 70.2.

(ii) Recall that L_k is the k-th Legendre polynomial, $L_k(1) = 1$, and $L_k(-1) = (-1)^k$; see §6.1. Then $\{(-1)^k L_k \circ T_n^{-1}, \{\frac{t-t_{n-1}}{t_{n,i}-t_{n-1}}\mathcal{L}_i \circ T_n^{-1}\}_{i\in\{1:k\}}\}$ are the Lagrange polynomials associated with the nodes $\{t_{n-1}, \{t_{n,i}\}_{i\in\{1:k\}}\}$. Since $u_{h\tau}$ is a member of $P_k^{\mathrm{g}}(\overline{J}_\tau;V_h)$, we have

$$u_{h\tau|J_n} = u_h^{n-1}(-1)^k L_k \circ T_n^{-1} + \sum_{i\in\{1:k\}} u_h^{n,i}\frac{t-t_{n-1}}{t_{n,i}-t_{n-1}}\mathcal{L}_i \circ T_n^{-1}.$$

The conclusion follows by evaluating the right-hand side at t_n since $L_k(1) = 1$ and $\frac{t_n-t_{n-1}}{t_{n,i}-t_{n-1}}\mathcal{L}_i(1) =: \alpha_i$ for all $i \in \{1{:}k\}$. □

70.1.4 Collocation schemes

We now briefly discuss a connection that exists between IRK schemes, dG(k) and cPG(k) schemes, and another class of methods called *collocation schemes*.

Definition 70.6 (Collocation). *Let $s \in \mathbb{N}\backslash\{0\}$. Let $\{\xi_l\}_{l\in\{1:s\}} \subset \widehat{J} := (-1,1]$ be s distinct numbers and set $t_{n,l} := T_n(\xi_l) \in J_n$ for all $l \in \{1{:}s\}$ and all $n \in \mathcal{N}_\tau$. A collocation scheme associated with the s points $\{\xi_l\}_{l\in\{1:s\}}$ for the time discretization of* (69.8) *seeks a function $\tilde{u}_{h\tau} \in P_s^{\mathrm{g}}(\overline{J}_\tau;V_h)$ as follows: First one sets $\tilde{u}_{h\tau}(0) := \mathcal{P}_{V_h}(u_0)$ and then for all $n \in \mathcal{N}_\tau$ one solves the following equations on $\{\tilde{u}_{h\tau}(t_{n,l})\}_{l\in\{1:s\}} \subset V_h$:*

$$\partial_t \tilde{u}_{h\tau}(t_{n,l}) + A_h(t_{n,l})(\tilde{u}_{h\tau}(t_{n,l})) = f_h(t_{n,l}), \qquad \forall l \in \{1{:}s\}. \tag{70.16}$$

Notice that since the $(s+1)$ numbers $\{-1, \{\xi_l\}_{l\in\{1:s\}}\}$ are distinct, $\tilde{u}_{h\tau|\overline{J}_n}$ is uniquely determined by the values it takes at these points for all $n \in \mathcal{N}_\tau$.

Proposition 70.7 (cPG(k) and dG(k) are collocation schemes). (i) *Let $k \geq 1$. Then $u_{h\tau} \in P_k^{\mathrm{g}}(\overline{J}_\tau;V_h)$ solves the cPG(k) scheme* (70.9) *if and only if $\tilde{u}_{h\tau} := u_{h\tau}$ solves the collocation scheme* (70.16) *associated with the $s := k$ Gauss–Legendre nodes.* (ii.a) *Let $k \geq 0$. If $u_{h\tau} \in P_k^{\mathrm{b}}(\overline{J}_\tau;V_h)$ solves the dG(k) scheme* (69.16)*, then $\tilde{u}_{h\tau} := \mathcal{R}_\tau(u_{h\tau}) \in P_{k+1}^{\mathrm{g}}(\overline{J}_\tau;V_h)$ solves the collocation scheme* (70.16) *associated with the $s := k+1$ Gauss–Radau nodes.* (ii.b) *Conversely if $\tilde{u}_{h\tau} \in \mathbb{P}_{k+1}^{\mathrm{g}}(\overline{J}_\tau;V_h)$ solves this collocation scheme, $u_{h\tau} := \mathcal{I}_k^{\mathrm{GR}}(\tilde{u}_{h\tau}) \in P_k^{\mathrm{b}}(\overline{J}_\tau;V_h)$ solves the dG(k) scheme* (69.16).

Proof. The assertion (i) follows from Proposition 70.2. The assertion (ii.a) follows from (69.20) (see Proposition 69.7) since $\mathcal{R}_\tau(u_{h\tau})(t_{n,l}) = u_{h\tau}(t_{n,l})$ for all $l \in \{1\!:\!k+1\}$ and all $n \in \mathcal{N}_\tau$, whereas the assertion (ii.b) follows from the same identity once we observe that $\mathcal{I}_k^{\mathrm{GR}}(\mathcal{R}_\tau(u_{h\tau})) = u_{h\tau}$. □

Combining the equivalence result of Proposition 70.7 with those from Lemma 69.11 (dG(k) ⇔ Radau IIA IRK) and Lemma 70.5 (cPG(k) ⇔ KB IRK) establishes that both IRK schemes are collocation methods. The connection between IRK schemes and collocation methods has been explored in Guillou and Soulé [172, p. 18] (see also Vlasák and Roskovec [282] for a related discussion on dG(k) schemes).

70.2 Stability and error analysis

In this section, we study the stability and the convergence properties of the cPG(k) scheme (70.8).

70.2.1 Stability

The choice of the spaces in (70.6) does not make it possible to take the discrete solution as the test function to prove a stability property for the bilinear form b_τ associated with the cPG(k) scheme. One must approximate in time the discrete solution with a polynomial of degree $(k-1)$. To this purpose, we use the Lagrange interpolation operator $\mathcal{I}_{k-1}^{\mathrm{GL}}$ associated with the Gauss–Legendre nodes and defined in (70.2). Let us equip $X_{h\tau}$ with the norm

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \|\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})\|_{L^2(J;V)}^2 + \frac{1}{2\alpha}\Big(\|v_{h\tau}(T)\|_L^2 + \|v_{h\tau}(0)\|_L^2\Big), \tag{70.17}$$

where $\alpha > 0$ is the coercivity constant of the bilinear form a. Notice that $\|\cdot\|_{X_{h\tau}}$ defines a norm on $X_{h\tau}$. To show that this is indeed the case, we use that $v_{h\tau|J_n} = \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau}) + \lambda_n \Phi_k$, for some $\lambda_n \in \mathbb{R}$ and $\Phi_{k|J_n} := L_k \circ T_n^{-1}$ for all $v_{h\tau} \in X_{h\tau}$ and all $n \in \mathcal{N}_\tau$, where L_k is the k-th Legendre polynomial on $\widehat{J}$. If $\|v_{h\tau}\|_{X_{h\tau}} = 0$, then $\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})_{|J_1} = 0$, and $v_{h\tau}(0) = 0$. Because $v_{h\tau|J_1} \in C^0(\overline{J}_1; V_h)$ and $L_k(-1) = (-1)^k \neq 0$, this implies that $\lambda_1 = 0$, i.e., $v_{h\tau|J_1} = 0$. We conclude that $v_{h\tau|J_n} = 0$ by induction on $n \in \mathcal{N}_\tau$.

Lemma 70.8 (Biased coercivity). (i) *The following holds true for all $v_{h\tau} \in X_{h\tau}$:*

$$b_\tau(v_{h\tau}, \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})) \geq \alpha \, \|v_{h\tau}\|_{X_{h\tau}}^2. \tag{70.18}$$

(ii) *The discrete problem* (70.8) *is well-posed.*

Proof. (i) Let $v_{h\tau} \in X_{h\tau}$. Using the coercivity of $a(t;\cdot,\cdot)$ at the k Gauss–Legendre nodes and the fact that the weights ω_l are all positive, we obtain

$$\begin{aligned}\int_{J_n} a(t; v_{h\tau}(t), \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(t)) \, \mu_k^{\mathrm{GL}}(\mathrm{d}t) &= \sum_{l\in\{1:k\}} \omega_{n,l} a(t_{n,l}; v_{h\tau}(t_{n,l}), v_{h\tau}(t_{n,l})) \\ &\geq \alpha \sum_{l\in\{1:k\}} \omega_{n,l} \|v_{h\tau}(t_{n,l})\|_V^2 = \alpha \int_{J_n} \|\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(t)\|_V^2 \mathrm{d}t,\end{aligned}$$

since the integrand is in $\mathbb{P}_{2k-2}(J_n;\mathbb{R})$ and the quadrature is of order $(2k-1)$. Moreover, evaluating the time integral using the Gauss–Legendre quadrature, we observe that

$$\int_{J_n} (\partial_t v_{h\tau}(t), \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(t))_L \mathrm{d}t = \int_{J_n} (\partial_t v_{h\tau}(t), v_{h\tau}(t))_L \mathrm{d}t,$$

since $\partial_t v_{h\tau}$ is a polynomial of degree $(k-1)$ at most and the quadrature is of order $(2k-1)$. Summing over $n \in \mathcal{N}_\tau$ we infer that

$$\int_J (\partial_t v_{h\tau}(t), \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(t))_L \mathrm{d}t = \frac{1}{2}\|v_{h\tau}(T)\|_L^2 - \frac{1}{2}\|v_{h\tau}(0)\|_L^2.$$

Using this identity, the lower bound just established above, and observing that $\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(0) := v_{h\tau}(0)$, we obtain

$$\begin{aligned} b_\tau(v_{h\tau}, \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})) \ge (v_{h\tau}(0), \mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})(0))_L + \frac{1}{2}\|v_{h\tau}(T)\|_L^2 - \frac{1}{2}\|v_{h\tau}(0)\|_L^2 \\ + \alpha\|\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})\|_{L^2(J;V)}^2 = \alpha\|v_{h\tau}\|_{X_{h\tau}}^2. \end{aligned}$$

(ii) The well-posedness of (70.8) follows from $\dim(X_{h\tau}) = \dim(Y_{h\tau})$ and the uniqueness of the solution implied by (70.18). □

70.2.2 Error analysis

Let $Z \in \{L, V_h\}$ and recall that $H^1(J;Z) \hookrightarrow C^0(\overline{J};Z)$. We proceed as in §69.3.2, but this time to handle the consistency error optimally it is convenient to introduce the approximation operator $\Pi_\tau^k : H^1(J;Z) \to P_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ defined as follows: For all $v \in H^1(J;Z)$,

$$\Pi_\tau^k(v)(0) = v(0), \tag{70.19a}$$

$$\int_J (\partial_t(\Pi_\tau^k(v) - v), \partial_t q_\tau)_L \mathrm{d}t = 0, \qquad \forall q_\tau \in P_k^{\mathrm{g}}(\overline{J}_\tau;Z). \tag{70.19b}$$

The above definition can be extended to $Z := V'$ by replacing the L-inner product in (70.19b) by the duality bracket between V' and V and by taking $q_\tau \in \mathbb{P}_k^{\mathrm{g}}(\overline{J}_\tau;V)$. Since Π_τ^k leaves $P_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ pointwise invariant and is $L^\infty(J;Z)$-stable uniformly w.r.t. τ, there is c s.t. for all $\tau > 0$ and all $v \in H^{k+1}(J;Z)$,

$$\|v - \Pi_\tau^k(v)\|_{L^2(J;Z)} \le c\,\tau^{k+1}|v|_{H^{k+1}(J;Z)}. \tag{70.20}$$

The definition of Π_τ^k is motivated by the following result.

Lemma 70.9 (Orthogonality). *The following identity holds true for all $v \in H^1(J;L)$ and all $y_\tau \in P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;L)$,*

$$\int_J (\partial_t(v - \Pi_\tau^k(v)), y_\tau)_L \mathrm{d}t = 0. \tag{70.21}$$

Proof. Let $y_\tau \in P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;L)$ and let $z_\tau \in P_k^{\mathrm{b}}(\overline{J}_\tau;L)$ be s.t. $z_\tau(0) := \mathcal{I}_{k-1}^{\mathrm{GL}}(y_\tau)(0) = y_\tau(0)$ and $z_\tau(t) := y_\tau(0) + \int_0^t y_\tau(s)\,\mathrm{d}s$ for all $t \in J$. By construction, we have $z_\tau \in P_k^{\mathrm{g}}(\overline{J}_\tau;L)$ and $\partial_t z_\tau = y_\tau$. As a result, we have $\int_J (\partial_t(v - \Pi_\tau^k(v)), y_\tau)_L \mathrm{d}t = \int_J (\partial_t(v - \Pi_\tau^k(v)), \partial_t z_\tau)_L \mathrm{d}t = 0$. □

Remark 70.10 (Other definition). Let $\Xi_{k-1}^{\mathrm{b}} : L^2(J;L) \to P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;L)$ be the $L^2(J;L)$-orthogonal projection. Then setting for all $v \in H^1(J;L)$, $\Pi_\tau^k(v)(t) = v(0) + \int_0^t \Xi_{k-1}^{\mathrm{b}}(\partial_t v)\,\mathrm{d}s$ for all $t \in J$, is equivalent to defining Π_τ^k using (70.19). □

To separate the time approximation and the space approximation, we assume that we have at hand a time-independent space approximation operator $\Pi_h \in \mathcal{L}(V;V_h)$ with $\|\Pi_h\|_{\mathcal{L}(V;V_h)}$ uniformly bounded w.r.t. $h \in \mathcal{H}$. For instance, we could take the quasi-interpolation operator $\mathcal{I}_h^{\mathrm{av}}$ or the elliptic projection Π_h^{e} if the bilinear form a is time-independent. We extend Π_h

to $\mathcal{L}(L^2(J;V);L^2(J;V_h))$ by setting $\Pi_h(v)(t) := \Pi_h(v(t))$ for all $v \in L^2(J;V)$. Notice that $\partial_t\Pi_h(v) = \Pi_h(\partial_t v)$ for all $v \in H^1(J;V)$ owing to Lemma 64.34. We are also going to use the commuting property $\Pi_\tau^k(\Pi_h(v)) = \Pi_h(\Pi_\tau^k(v))$ for all $v \in H^1(J;V)$; see Exercise 70.5.

We extend the $\|\cdot\|_{X_{h\tau}}$-norm defined in (70.17) to $H^1(J;V)$ (we use the same notation for simplicity). Recalling the time scale $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$, where $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, we set $\rho_\sharp := \max(\rho, T)$.

Theorem 70.11 ($L^2(J;V)$-estimate). *Let $u \in X$ solve (66.3) and $u_{h\tau} \in X_{h\tau}$ solve (70.8). Assume that $u \in H^{k+2}(J;V') \cap W^{k+1,\infty}(J;V)$. Let $c_1(u) := \frac{1}{\alpha}|u|_{H^{k+2}(J;V')} + \xi_\kappa T^{\frac12}|u|_{W^{k+1,\infty}(J;V)}$ with the contrast factor $\xi_\kappa := \frac{M}{\alpha}$. There is c s.t. for all $h \in \mathcal{H}$, α, M, and T,*

$$\|u-u_h\|_{X_{h\tau}} \le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0-\Pi_h(u_0)\|_L + c\,\Big(\tau^{k+1}c_1(u) + \frac{1}{\alpha}\|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + \xi_\kappa\rho_\sharp^{\frac12}\|u-\Pi_h(u)\|_{L^\infty(J;V)}\Big). \tag{70.22}$$

Proof. Let $y_{h\tau} \in Y_{h\tau} := P_{k-1}^{\mathrm{b}}(\overline{J}_\tau;V_h)$. We have

$$\begin{aligned} b_\tau(u_{h\tau},y_{h\tau}) &= (u_0,y_{h\tau}(0))_L + \int_J \langle f,y_{h\tau}\rangle_{V',V}\mu_k^{\mathrm{GL}}(\mathrm{d}t)\\ &= (u_0,y_{h\tau}(0))_L + R(y_{h\tau}) + \int_J(\partial_t u,y_{h\tau})_L\mathrm{d}t + \int_J a(t;u,y_{h\tau})\mu_k^{\mathrm{GL}}(\mathrm{d}t), \end{aligned}$$

with $R(y_{h\tau}) := \int_J(\partial_t u,y_{h\tau})_L\mu_k^{\mathrm{GL}}(\mathrm{d}t) - \int_J(\partial_t u,y_{h\tau})_L\mathrm{d}t$ and where we used that $(\partial_t u(t_{n,l}),y_{h\tau})_L + a(t_{n,l};u(t_{n,l}),y_{h\tau}) = \langle f(t_{n,l}),y_{h\tau}\rangle_{V',V}$ for all $n \in \mathcal{N}_\tau$ and all $l \in \{1{:}k\}$ owing to the regularity assumption on u. Let us introduce $v_{h\tau} := \Pi_\tau^k(\Pi_h(u)) \in X_{h\tau} := P_k^{\mathrm{g}}(\overline{J}_\tau;V_h)$ and let us set $e_{h\tau} := u_{h\tau} - \Pi_\tau^k(\Pi_h(u))$ and $\eta := u - \Pi_\tau^k(\Pi_h(u))$. Using the above calculation for $b_\tau(u_{h\tau},y_{h\tau})$, we obtain

$$\begin{aligned} b_\tau(u_{h\tau}-v_{h\tau},y_{h\tau}) &= (u_0-v_{h\tau}(0),y_{h\tau}(0))_L + R(y_{h\tau})\\ &\quad + \int_J(\partial_t(u-v_{h\tau}),y_{h\tau})_L\mathrm{d}t + \int_J a(t;u-v_{h\tau},y_{h\tau})\mu_k^{\mathrm{GL}}(\mathrm{d}t). \end{aligned}$$

By definition of Π_τ^k, we have $v_{h\tau}(0) = \Pi_h(u_0)$. Moreover, using Lemma 70.9 for the function $v := \Pi_h(u)$ and since $\partial_t(\Pi_h(u)) = \Pi_h(\partial_t u)$, we infer that

$$\begin{aligned} b_\tau(e_{h\tau},y_{h\tau}) &= (\eta(0),y_{h\tau}(0))_L + R(y_{h\tau}) + \int_J(\partial_t u - \Pi_h(\partial_t u),y_{h\tau})_L\mathrm{d}t\\ &\quad + \int_J a(t;\eta,y_{h\tau})\mu_k^{\mathrm{GL}}(\mathrm{d}t). \end{aligned}$$

Let $\mathfrak{T}_1,\dots,\mathfrak{T}_4$ denote the four terms on the right-hand side. We have $|\mathfrak{T}_1| \le \|u_0-\Pi_h(u_0)\|_L\|y_{h\tau}(0)\|_L$. Since (70.5) implies that $\int_J(\partial_t u,y_{h\tau})_L\mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_J(\mathcal{I}_k^{\mathrm{GL}+}(\partial_t u),y_{h\tau})_L\mathrm{d}t$, we infer that $R(y_{h\tau}) = \int_J(\mathcal{I}_k^{\mathrm{GL}+}(\partial_t u)-\partial_t u,y_{h\tau})_L\mathrm{d}t$. The approximation property (70.4) of $\mathcal{I}_k^{\mathrm{GL}+}$ gives

$$|\mathfrak{T}_2| \le c\,\tau^{k+1}|u|_{H^{k+2}(J;V')}\|y_{h\tau}\|_{L^2(J;V)}.$$

Moreover, we have $|\mathfrak{T}_3| \le \|\partial_t u-\Pi_h(\partial_t u)\|_{L^2(J;V')}\|y_{h\tau}\|_{L^2(J;V)}$. Finally, since Π_τ^k and Π_h commute, the stability of Π_h and the approximation property (70.20) of Π_τ^k imply that for all $l \in \{1{:}k\}$ and all $n \in \mathcal{N}_\tau$,

$$\begin{aligned} \|\eta(t_{n,l})\|_V &\le \|(u-\Pi_h(u))(t_{n,l})\|_V + \|\Pi_h\|_{\mathcal{L}(V;V_h)}\|(u-\Pi_\tau^k(u))(t_{n,l})\|_V\\ &\le \|u-\Pi_h(u)\|_{L^\infty(J;V)} + c\,\tau^{k+1}|u|_{W^{k+1,\infty}(J;V)} =: C(u). \end{aligned}$$

Invoking the boundedness of a, the Cauchy–Schwarz inequality, and since the quadrature is of order $(2k-1)$ and $\sum_{n\in\mathcal{N}_\tau}\sum_{l\in\{1:k\}}\omega_{n,l}=T$, we infer that

$$|\mathfrak{T}_4| \le \sum_{n\in\mathcal{N}_\tau}\sum_{l\in\{1:k\}} \omega_{n,l} M C(u) \|y_{h\tau}(t_{n,l})\|_V \le M T^{\frac12} C(u) \|y_{h\tau}\|_{L^2(J;V)}.$$

Let us set $\|y_{h\tau}\|_{Y_{h\tau}}^2 := \|y_{h\tau}\|_{L^2(J;V)}^2 + \frac{1}{\alpha}\|y_{h\tau}(0)\|_L^2$. Combining the above estimates and recalling the definition of $c_1(u)$ in the assertion shows that

$$\begin{aligned}\sup_{y_{h\tau}\in X_{h\tau}} \frac{|b(e_{h\tau},y_{h\tau})|}{\|y_{h\tau}\|_{Y_{h\tau}}} &\le \sqrt{2\alpha}\|u_0-\Pi_h(u_0)\|_L + c\,\alpha\tau^{k+1}c_1(u)\\ &\quad + \|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + MT^{\frac12}\|u-\Pi_h(u)\|_{L^\infty(J;V)}.\end{aligned}$$

We now invoke the biased coercivity property (70.18) which gives

$$\alpha\,\|e_{h\tau}\|_{X_{h\tau}} \le \frac{b_\tau(e_{h\tau},\mathcal{I}_{k-1}^{\mathrm{GL}}(e_{h\tau}))}{\|e_{h\tau}\|_{X_{h\tau}}} \le c \sup_{y_{h\tau}\in X_{h\tau}} \frac{|b_\tau(e_{h\tau},y_{h\tau})|}{\|y_{h\tau}\|_{Y_{h\tau}}},$$

where we used that $\|\mathcal{I}_{k-1}^{\mathrm{GL}}(e_{h\tau})\|_{Y_{h\tau}} \le \|e_{h\tau}\|_{X_{h\tau}}$ since $\mathcal{I}_{k-1}^{\mathrm{GL}}(e_{h\tau})(0)=e_{h\tau}(0)$. Combining the above two bounds and using the definition of ξ_κ yields

$$\begin{aligned}\|e_{h\tau}\|_{X_{h\tau}} &\le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0-\Pi_h(u_0)\|_L + c\,\tau^{k+1}c_1(u)\\ &\quad + \frac{1}{\alpha}\|\partial_t u - \Pi_h(\partial_t u)\|_{L^2(J;V')} + \xi_\kappa T^{\frac12}\|u-\Pi_h(u)\|_{L^\infty(J;V)}.\end{aligned}$$

Finally, the triangle inequality implies that $\|u-u_h\|_{X_{h\tau}} \le \|e_{h\tau}\|_{X_{h\tau}} + \|\eta\|_{X_{h\tau}}$. Using the definition of the time scales ρ and $\rho_\sharp$ yields

$$\|\eta\|_{X_{h\tau}} \le \|\eta\|_{L^2(J;V)} + c\,\frac{\iota_{L,V}}{\sqrt{\alpha}}\|\eta\|_{L^\infty(J;V)} \le c'\,\rho_\sharp^{\frac12}\|\eta\|_{L^\infty(J;V)}.$$

Reasoning as above then shows that $\|\eta\|_{L^\infty(J;V)} \le \|u-\Pi_h(u)\|_{L^\infty(J;V)} + c\,\tau^{k+1}|u|_{W^{k+1,\infty}(J;V)}$. Putting everything together concludes the proof. □

Remark 70.12 (Optimality in time). The identity (70.21) satisfied by the operator Π_τ^k and using the interpolation operator $\mathcal{I}_k^{\mathrm{GL}+}$ are the two key ideas to achieve an optimal error estimate in time. □

Remark 70.13 (Supercloseness). Assume that a is time-independent so that one can use the elliptic projector $\Pi_h := \Pi_h^{\mathrm{E}}$ in the proof of Theorem 70.11. Arguing as in Remark 69.20 for dG(k) schemes gives

$$\|e_{h\tau}\|_{X_{h\tau}} \le \frac{\sqrt{2}}{\sqrt{\alpha}}\|u_0-\Pi_h^{\mathrm{E}}(u_0)\|_L + c\,\tau^{k+1}c_1(u) + \frac{1}{\alpha}\|\partial_t u - \Pi_h^{\mathrm{E}}(\partial_t u)\|_{L^2(J;V')}.$$ □

Remark 70.14 (Convergence, heat equation). Let us consider the approximation of the heat equation with H^1-conforming finite elements. Let $r\in[1,k']$, where $k'\ge 1$ is the degree of the finite elements used to build V_h. Assume that $u\in H^{k+2}(J;H^{-1}(D))\cap W^{k+1,\infty}(J;H_0^1(D))\cap W^{1,\infty}(J;H^{k'+1}(D))$. Then the estimate from Theorem 70.11 implies that $\|u-u_{h\tau}\|_{L^2(J;H_0^1)}$ decays as $\mathcal{O}(\tau^{k+1}c_1(u)+h^{k'}c_2(u))$ with $c_2(u) := \rho_\sharp^{\frac12}\|u\|_{W^{1,\infty}(J;H^{k'+1})}$. Moreover, the estimate

from Remark 70.13 implies that $\|(u-u_{h\tau})(T)\|_{L^2(D)}$ decays as $\mathcal{O}(\tau^{k+1}c_1(u)+h^{k'+s}\ell_D^{-s}c_2(u))$, where $s\in(0,1]$ is the elliptic regularity pickup index ($s=1$ if there is full elliptic regularity pickup). Finally, since the constant c in the estimate does not depend on T, the error $\sup_{n\in\mathcal{N}_\tau}\|(u-u_{h\tau})(t_n)\|_{L^2(D)}$ decays with the same rate. □

Remark 70.15 (Literature). Further developments on the error analysis can be found in Aziz and Monk [19]. In particular, [19, Thm. 3.4] shows for the heat equation with full elliptic regularity that $\|u-u_{h\tau}\|_{L^\infty(\overline{J};L)}\le c\big(\tau^{k+1}(\rho|u|_{H^{k+2}(J;L)}+\sqrt{\alpha\rho}|u|_{W^{k+1,\infty}(J;V)})+\|u-\Pi_h^{\mathrm{E}}(u)\|_{L^2(J;L)}\big)$. Under more restrictive smoothness assumptions, [19, Thm. 4.2] establishes that the error in time decays as $\mathcal{O}(\tau^{2k})$ for $k\ge1$. □

70.3 Algebraic realization

Let $\{t_{n,l}\}_{l\in\{1:k\}}$ be the Gauss–Legendre nodes in the time interval J_n for all $n\in\mathcal{N}_\tau$. Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h, e.g., the global shape functions in the finite element space V_h. Recall the mass matrix $\mathcal{M}\in\mathbb{R}^{I\times I}$, the time-dependent stiffness matrix $\mathcal{A}^{n,p}\in\mathbb{R}^{I\times I}$, and the load vectors $\mathsf{F}^{n,p}\in\mathbb{R}^I$ defined in (69.31) for all $n\in\mathcal{N}_\tau$ and all $p\in\{1{:}k\}$, that is, $\mathcal{M}_{ij}:=(\varphi_j,\varphi_i)_L$, $\mathcal{A}_{ij}^{n,p}:=a(t_{n,p};\varphi_j,\varphi_i)$, and $\mathsf{F}_i^{n,p}:=\langle f(t_{n,p}),\varphi_i\rangle_{V',V}$ for all $i,j\in\{1{:}I\}$.

70.3.1 IRK implementation

Since the solution produced by the cPG(k) scheme and the KB IRK scheme are identical according to Lemma 70.5, one way to implement the method is to use the IRK strategy (70.12)-(70.14) with $s:=k$ stages. One first sets $\mathsf{U}^0\in\mathbb{R}^I$ so that $\mathcal{P}_{V_h}(u_0)=\sum_{i\in\{1:I\}}\mathsf{U}_i^0\varphi_i$. Then for every $n\ge1$, letting $\mathsf{U}^{n,p}\in\mathbb{R}^I$ be the coordinate vector of $u_h^{n,p}$ for all $p\in\{1{:}k\}$, (70.12) amounts to solving the following linear system with $\{a_{pq}\}_{p,q\in\{1:k\}}$ given in (70.11):

$$\begin{pmatrix}\mathcal{M}+\tau a_{11}\mathcal{A}^{n,1} & \cdots & \tau a_{1k}\mathcal{A}^{n,k}\\ \vdots & \ddots & \vdots\\ \tau a_{m1}\mathcal{A}^{n,1} & \cdots & \mathcal{M}+\tau a_{kk}\mathcal{A}^{n,k}\end{pmatrix}\begin{pmatrix}\mathsf{U}^{n,1}\\ \vdots\\ \mathsf{U}^{n,k}\end{pmatrix}=\begin{pmatrix}\widetilde{\mathsf{G}}^{n,1}\\ \vdots\\ \widetilde{\mathsf{G}}^{n,k}\end{pmatrix},\tag{70.23}$$

nd the load vectors defined by $\widetilde{\mathsf{G}}^{n,p}:=\mathcal{M}u_h^{n-1}+\tau\sum_{q\in\{1:k\}}a_{pq}\mathsf{F}^{n,q}\in\mathbb{R}^I$ for all $p\in\{1{:}k\}$. Finally, one sets $\mathsf{U}^n:=\alpha_0\mathsf{U}^{n-1}+\sum_{p\in\{1:k\}}\alpha_p\mathsf{U}^{n,p}$ with $\{\alpha_p\}_{p\in\{1:k\}}$ defined in (70.14).

70.3.2 General case

We now consider the general case and write the linear system corresponding to the cPG(k) formulation (70.9) for general bases of $\mathbb{P}_k(\widehat{J};\mathbb{R})$ and $\mathbb{P}_{k-1}(\widehat{J};\mathbb{R})$. Let $\{\phi_q\}_{q\in\{0:k\}}$ be a basis of $\mathbb{P}_k(\widehat{J};\mathbb{R})$ and $\{\psi_p\}_{p\in\{1:k\}}$ a basis of $\mathbb{P}_{k-1}(\widehat{J};\mathbb{R})$. For simplicity, we assume that

$$\phi_0(-1)=1,\quad \phi_0(\xi_l)=0,\quad\text{and}\quad \phi_l(-1)=0,\tag{70.24}$$

for all $l\in\{1{:}k\}$, where $\{\xi_l\}_{l\in\{1:k\}}$ are the Gauss–Legendre nodes in $\widehat{J}$. This implies $\phi_0(t)=(-1)^kL_k(t)$; see §6.1.

For all $n\in\mathcal{N}_\tau$ and all $q\in\{0{:}k\}$, we introduce the coordinate vectors $\mathsf{U}^{n,q}\in\mathbb{R}^I$ s.t. $u_{h\tau}(\boldsymbol{x},t):=\sum_{j\in\{1:I\}}\sum_{q\in\{0:k\}}\mathsf{U}_j^{n,q}\phi_q(T_n^{-1}(t))\varphi_j(\boldsymbol{x})$ for all $(\boldsymbol{x},t)\in D\times J_n$. For all $n\in\overline{\mathcal{N}}_\tau$, we also introduce

$\mathsf{U}^n \in \mathbb{R}^I$ so that $u_{h\tau}(\boldsymbol{x}, t_n) = u_h^n(\boldsymbol{x}) := \sum_{j\in\{1:I\}} \mathsf{U}_j^n \varphi(\boldsymbol{x})$. The constraints (70.24) and the identity $u_{h\tau}(\boldsymbol{x}, t_{n-1}) = u_h^{n-1}$ imply that $\mathsf{U}^{n,0} = \mathsf{U}^{n-1}$. Thus, at the beginning of the n-th step in (70.9), $\mathsf{U}^{n,0}$ is known either from the initial condition ($n = 1$) or the previous time-step ($n \ge 2$).

For every integers $p, q \in \{1{:}k\}$, we introduce the coefficients

$$b_{pq} := \int_{-1}^{1} \phi_q'(s)\psi_p(s)\,\mathrm{d}s, \qquad d_p := -\int_{-1}^{1} \phi_0'(s)\psi_p(s)\,\mathrm{d}s. \tag{70.25}$$

Then, considering test functions of the form $\varphi_i(\boldsymbol{x})\psi_p(T_n^{-1}(t))$ for all $i \in \{1{:}I\}$ and all $p \in \{1{:}k\}$, we rewrite the cPG(k) scheme (70.9) in the following block form:

$$\begin{pmatrix} b_{11}\mathcal{M} + \tau\mathcal{A}^{n,11} & \cdots & b_{1k}\mathcal{M} + \tau\mathcal{A}^{n,1k} \\ \vdots & \ddots & \vdots \\ b_{k1}\mathcal{M} + \tau\mathcal{A}^{n,k1} & \cdots & b_{kk}\mathcal{M} + \tau\mathcal{A}^{n,kk} \end{pmatrix} \begin{pmatrix} \mathsf{U}^{n,1} \\ \vdots \\ \mathsf{U}^{n,k} \end{pmatrix} = \begin{pmatrix} \mathsf{G}^{n,1} \\ \vdots \\ \mathsf{G}^{n,k} \end{pmatrix},$$

with the stiffness matrices $\mathcal{A}^{n,pq} := \sum_{l\in\{1:k\}} \frac{\omega_l}{2}\psi_q(\xi_l)\psi_p(\xi_l)\mathcal{A}_l^n \in \mathbb{R}^{I\times I}$ for all $p, q \in \{1{:}k\}$ and all $n \in \mathcal{N}_\tau$, and the load vectors $\mathsf{G}^{n,p} := d_p\mathcal{M}\mathsf{U}^{n-1} + \tau\sum_{q\in\{1:k\}} \frac{\omega_q}{2}\psi_p(\xi_q)\mathsf{F}^{n,q} \in \mathbb{R}^I$ for all $p \in \{1{:}k\}$ and all $n \in \mathcal{N}_\tau$. To prepare for the next time-step, we finally set $\mathsf{U}^n := \sum_{q\in\{0:k\}} \alpha_q \mathsf{U}^{n,q}$ with $\alpha_q := \phi_q(1)$ for all $q \in \{0{:}k\}$.

The cPG(k) scheme is only slightly less expensive than the dG(k) scheme since for all $n \in \mathcal{N}_\tau$, it requires assembling k stiffness matrices (instead of $(k+1)$) and solving a globally coupled linear system of size $I\times k$ (instead of $I\times(k+1)$). The global system matrix is nonsymmetric for both schemes even if the bilinear form a is symmetric. If the bilinear form a is time-independent, the assembling of the global system matrix is simplified since we have $\mathcal{A}^{n,pq} = m_{pq}\mathcal{A}$ with the time-independent stiffness matrix $\mathcal{A} \in \mathbb{R}^{I\times I}$ s.t. $\mathcal{A}_{ij} := a(\varphi_j, \varphi_i)$ for all $i, j \in \{1{:}I\}$ (see (69.36)), and the coefficients m_{pq} s.t. $m_{pq} := \frac{1}{2}\int_{-1}^{1} \phi_q(s)\psi_p(s)\,\mathrm{d}s$ for all $p, q \in \{1{:}k\}$. Hence, it is only necessary to assemble *one* stiffness matrix. As for the dG(k) scheme, the algebraic formulation of the cPG(k) scheme can be rewritten in a more compact form using tensor notation as follows:

$$(\mathcal{M}\otimes\mathbb{B} + \tau\mathcal{A}\otimes\mathbb{M})\mathbf{U}^n = \mathbf{G}^n, \tag{70.26}$$

with $\mathbf{U}^n := (\mathsf{U}^{n,1}, \dots, \mathsf{U}^{n,k})^\mathsf{T} \in \mathbb{R}^{Ik}$ and $\mathbf{G}^n := (\mathsf{G}^{n,1}, \dots, \mathsf{G}^{n,k})^\mathsf{T} \in \mathbb{R}^{Ik}$. The linear system (70.26) can be symmetrized and preconditioned by proceeding as in Remark 69.23.

Remark 70.16 (Diagonal $\mathbb{M}$). One can choose for $\{\phi_q\}_{q\in\{0:k\}}$ the Lagrange interpolation polynomials associated with the nodes $\{-1, \xi_1, \dots, \xi_k\}$ and for $\{\psi_p\}_{p\in\{1:k\}}$ the Lagrange interpolation polynomials associated with the nodes $\{\xi_p\}_{p\in\{1:k\}}$. This choice is compatible with the assumption (70.24). One advantage of this choice is that the matrix $\mathbb{M}$ becomes diagonal, $m_{pq} = \delta_{pq}\frac{\omega_p}{2}$, and the load term becomes $\mathsf{G}^{n,p} := d_p\mathcal{M}\mathsf{U}^{n-1} + \tau\frac{\omega_p}{2}\mathsf{F}^{n,p}$. We also have $d_p = -\omega_p\phi_0'(\xi_p)$. See Exercise 70.4 for $k := 1$ (Crank–Nicolson) and $k := 2$ (see also Hussain et al. [193]). □

Exercises

Exercise 70.1 (Interpolation operators). (i) Let $\mathcal{I}_{k-1}^{\mathrm{GL}}$ be the Lagrange interpolation operator defined in (70.2) using $Z := L$. Prove that

$$\int_J (p, \mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L \mathrm{d}t = \int_J (p, w)_L\, \mu_k^{\mathrm{GL}}(\mathrm{d}t), \tag{70.27a}$$

$$\int_J (v, \mathcal{I}_{k-1}^{\mathrm{GL}}(w))_L\, \mu_k^{\mathrm{GL}}(\mathrm{d}t) = \int_J (v, w)_L\, \mu_k^{\mathrm{GL}}(\mathrm{d}t), \tag{70.27b}$$

for all $p \in P_k^{\mathrm{b}}(J_\tau;L)$ and all $v,w \in L^2(J;L)$. (ii) Let $Z \subseteq L$. Prove that the restriction of $\mathcal{I}_{k-1}^{\mathrm{GL}}$ to $P_k^{\mathrm{g}}(\overline{J}_\tau;Z)$ coincides with the $L^2(J;Z)$-orthogonal projection onto $P_{k-1}^{\mathrm{b}}(J_\tau;Z)$. (iii) Prove (70.5).

Exercise 70.2 (Equivalence with KB IRK). Prove the converse assertion in Lemma 70.5. $\big($*Hint*: show that $u_{h\tau}(t) = u_h^{n-1} + \tau\sum_{j\in\{1:k\}} \frac{1}{2}\int_{-1}^{T_n^{-1}(t)} \mathcal{L}_j(\xi)\,\mathrm{d}\xi\big(f_h(t_{n,j}) - A_h(t_{n,j})(u_h^{n,j})\big)$ for all $t \in J_n$ and all $n \in \mathcal{N}_\tau$.$\big)$

Exercise 70.3 (Butcher simplifying assumptions). Let $s \in \mathbb{N}\setminus\{0\}$ and let $\{c_i\}_{i\in\{1:s\}}$ be s distinct points in $[0,1]$. Let $\xi_i := 2c_i - 1$ and $\mathcal{L}_i(\xi) := \prod_{j\in\{1:s\}\setminus\{i\}} \frac{\xi-\xi_j}{\xi_i-\xi_j}$ for all $i \in \{1:s\}$. Let $a_{ij} := \frac{1}{2}\int_{-1}^{2c_i-1} \mathcal{L}_j(\xi)\,\mathrm{d}\xi$, $b_i := \frac{1}{2}\int_{-1}^{1} \mathcal{L}_i(\xi)\,\mathrm{d}\xi$ for all $i \in \{1:s\}$. (i) Show that the set $\{\xi_i, 2b_i\}_{i\in\{1:s\}}$ is a quadrature of order $k_{\mathcal{Q}} \geq s-1$ over the interval $[-1,1]$ (see Definition 6.4). (*Hint*: observe that $p = \sum_{i\in\{1:s\}} p(\xi_i)\mathcal{L}_i$ for all $p \in \mathbb{P}_{s-1}(\widehat{J};\mathbb{R})$.) (ii) Show that for all $q \in \{1:s\}$,

$$\sum_{j\in\{1:s\}} a_{ij}c_j^{q-1} = \frac{c_i^q}{q},\ \forall i \in \{1:s\}, \qquad \sum_{j\in\{1:s\}} b_j c_j^{q-1} = \frac{1}{q}.$$

(*Hint*: integrate $\big(\frac{1+\xi}{2}\big)^{q-1}$ over $(-1,\xi_i)$ for all $i \in \{1:s\}$ and over $(-1,1)$.) (iii) Assuming that $k_{\mathcal{Q}} \geq s$, show that for all $j \in \{1:s\}$,

$$\sum_{i\in\{1:s\}} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1-c_j^q), \qquad \forall q \in \{1:k_{\mathcal{Q}}-s+1\}.$$

(*Hint*: integrate the polynomial $\big(\frac{1+\xi}{2}\big)^{q-1}\int_{-1}^{\xi} \mathcal{L}_j(\xi)\,\mathrm{d}\xi$ over $(-1,1)$.) *Note*: these formulae are called *Butcher's simplifying assumptions* in the ODE literature (see Butcher [77, Thm. 7], Hairer et al. [176, §II.6], [175, §IV.5, Thm. 5.1], see also the order conditions stated in Theorem 78.5).

Exercise 70.4 (cPG(k)). Assume that a is time-independent. (i) Use the IRK formalism and the tableaux in (70.15) to write the algebraic form of cPG(1) and cPG(2). (*Hint*: use the coefficients $\{\alpha_i\}_{i\in\{0:s\}}$.) (ii) Write again the cPG(1) and cPG(2) schemes in algebraic form using the formalism described in §70.3.2 and the bases from Remark 70.16. (*Hint*: for $k := 1$, it is of the form $(2\mathcal{M}+\tau\mathcal{A})\mathsf{U}^{n,1} = 2\mathcal{M}\mathsf{U}^{n-1} + \tau\mathsf{F}^{n,1}$ and $\mathsf{U}^n = 2\mathsf{U}^{n,1} - \mathsf{U}^{n-1}$, whereas for $k := 2$, it is of the form

$$\begin{pmatrix} \frac{3}{2} & \frac{2\sqrt{3}-3}{2} \\ -\frac{2\sqrt{3}+3}{2} & \frac{3}{2} \end{pmatrix}\begin{pmatrix} \mathcal{M}\mathsf{U}^{n,1} \\ \mathcal{M}\mathsf{U}^{n,2} \end{pmatrix} + \frac{\tau}{2}\begin{pmatrix} \mathcal{A}\mathsf{U}^{n,1} \\ \mathcal{A}\mathsf{U}^{n,2} \end{pmatrix} = \begin{pmatrix} \sqrt{3}\mathcal{M}\mathsf{U}^{n-1} + \frac{\tau}{2}\mathsf{F}^{n,1} \\ -\sqrt{3}\mathcal{M}\mathsf{U}^{n-1} + \frac{\tau}{2}\mathsf{F}^{n,2} \end{pmatrix},$$

and $\mathsf{U}^n := \mathsf{U}^{n-1} - \sqrt{3}(\mathsf{U}^{n,1} - \mathsf{U}^{n,2})$.)

Exercise 70.5 (Π_τ^k and Π_h commute). Let $\Pi_h \in \mathcal{L}(V;V_h)$. Show that $\Pi_\tau^k(\Pi_h(v)) = \Pi_h(\Pi_\tau^k(v))$ for all $v \in H^1(J;V)$. (*Hint*: use Remark 70.10 and prove that Π_h commutes with Ξ_{k-1}^{b} by introducing $(\Pi_h)^* \in \mathcal{L}(V_h;V')$.)

Chapter 71

Analysis using inf-sup stability

In this chapter, we revisit the well-posedness of the model parabolic problem (65.10), i.e., we give another proof of Lions' theorem (Theorem 65.9) using the framework of the BNB theorem (Theorem 25.9). In other words, we establish the well-posedness by proving an inf-sup condition. Then we exploit the inf-sup condition to revisit the stability and the error analysis for various approximation techniques investigated in the previous chapters: (1) the space semi-discrete problem considered in §66.2; (2) the implicit Euler scheme introduced in §67.1; (3) the dG(k) scheme investigated in Chapter 69; (4) the cPG(k) scheme investigated in Chapter 70.

71.1 Well-posedness

The goal of this section is to give another proof of Lions' theorem by using the BNB theorem.

71.1.1 Functional setting

Let $(V, L \equiv L', V')$ be a Gelfand triple and recall the functional spaces

$$X := X(J;V,V') = \{v \in L^2(J;V) \mid \partial_t v \in L^2(J;V')\}, \tag{71.1a}$$

$$Y := Y_0 \times Y_1, \qquad Y_0 := L, \qquad Y_1 := L^2(J;V). \tag{71.1b}$$

Let $A : J \to \mathcal{L}(V;V')$ be a linear operator that satisfies the properties (65.5). Let $\alpha(t)$ and $M(t)$ denote the coercivity and boundedness constants of $A(t) \in \mathcal{L}(V;V')$ for a.e. $t \in J$. The real numbers α and M introduced in the assumptions (65.5b)-(65.5c) are then $\alpha := \operatorname{ess\,inf}_{t\in J} \alpha(t) > 0$ and $M := \operatorname{ess\,sup}_{t\in J} M(t) < \infty$.

The model problem we consider in this chapter is (65.10), i.e.,

$$\text{We seek } u \in X \text{ s.t. } b(u,y) = \ell(y) \text{ for all } y := (y_0, y_1) \in Y, \tag{71.2}$$

with the forms $b : X \times Y \to \mathbb{R}$ and $\ell : Y \to \mathbb{R}$ s.t.

$$b(v,y) := (v(0), y_0)_L + \int_J \langle \partial_t v(t) + A(v)(t), y_1(t)\rangle_{V',V} \mathrm{d}t, \tag{71.3a}$$

$$\ell(y) := (u_0, y_0)_L + \int_J \langle f(t), y_1(t)\rangle_{V',V} \mathrm{d}t. \tag{71.3b}$$

Since we do not assume that A takes self-adjoint values, we denote by $M_{\mathrm{s}}(t)$ the boundedness constant of $A_{\mathrm{s}}(t) := \frac{1}{2}(A(t) + A(t)^*)$ and we set $M_{\mathrm{s}} := \operatorname{ess\,sup}_{t\in J} M_{\mathrm{s}}(t)$. We also need to consider the coercivity constant of $A(t)^{-1}$. Since the operator $A(t) \in \mathcal{L}(V;V')$ is coercive for a.e. $t \in J$, its inverse $A(t)^{-1} \in \mathcal{L}(V';V)$ is also coercive (see (C.29)). Let $\gamma(t)$ be the coercivity constant of $A(t)^{-1}$. If A takes self-adjoint values, we have $\gamma(t) = \frac{1}{M(t)}$ owing to Lemma C.63. In the general situation, Lemma C.64 shows that $\gamma(t) \in [\frac{\alpha(t)}{M(t)^2}, \frac{1}{M_{\mathrm{s}}(t)}]$ for a.e. $t \in J$. We then have

$$\gamma := \operatorname*{ess\,inf}_{t\in J} \gamma(t) \in [\tfrac{\alpha}{M^2}, \tfrac{1}{M_{\mathrm{s}}}]. \tag{71.4}$$

In what follows, we will use that $\gamma\alpha \le \gamma M_{\mathrm{s}} \le 1$, $\alpha \le M_{\mathrm{s}} \le M$, and

$$\langle \phi, A(t)^{-1}(\phi)\rangle_{V',V} \ge \gamma \, \|\phi\|_{V'}^2, \qquad \forall \phi \in V', \text{ for a.e. } t \in J. \tag{71.5}$$

We equip the spaces X and Y defined in (71.1) with the following norms:

$$\|v\|_X^2 := \|v\|_{L^2(J;V)}^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \frac{1}{\alpha}\|v(T)\|_L^2, \tag{71.6a}$$

$$\|y\|_Y^2 := \frac{1}{\alpha}\|y_0\|_L^2 + \|y_1\|_{L^2(J;V)}^2. \tag{71.6b}$$

The last term in (71.6a) is legitimate owing to the continuous embedding $X \hookrightarrow C^0(\overline{J};L)$ from Lemma 64.40. Other choices for the X-norm are possible (see Exercise 71.2). We also notice that the norms in X and Y are dimensionally consistent. The present choice appears to deliver relatively sharp bounds on the inf-sup and boundedness constants of the bilinear form b.

71.1.2 Boundedness and inf-sup stability

Let us start with the boundedness of the bilinear form b.

Lemma 71.1 (Boundedness). *Let us set $\theta := \frac{1}{2}\operatorname{ess\,sup}_{t\in J}\|C(t)\|_{\mathcal{L}(V)}$ with $C(t) := A(t)^{-*}A(t) - I_V \in \mathcal{L}(V)$. The following holds true:*

$$\sup_{v\in X}\sup_{y\in Y}\frac{|b(v,y)|}{\|v\|_X\|y\|_Y} \le M_b := \frac{(1+\theta)^{\frac{1}{2}}}{\gamma}. \tag{71.7}$$

Proof. Let $(v,y) \in X\times Y$. The Cauchy–Schwarz inequality implies that

$$|b(v,y)| \le \big(\|\partial_t v + A(v)\|_{L^2(J;V')}^2 + \alpha\|v(0)\|_L^2\big)^{\frac{1}{2}}\|y\|_Y.$$

Using the coercivity of $A(t)^{-1}$, rearranging the terms, and dropping the time dependency in the integrals to simplify the notation, we infer that

$$\begin{aligned}\|\partial_t v + A(v)\|_{L^2(J;V')}^2 &\le \frac{1}{\gamma}\int_J \langle \partial_t v + A(v), A^{-1}(\partial_t v) + v\rangle_{V',V}\mathrm{d}t\\ &= \frac{1}{\gamma}\int_J \Big(\langle \partial_t v, A^{-1}(\partial_t v)\rangle_{V',V} + \langle A(v), v\rangle_{V',V} + 2\langle \partial_t v, v\rangle_{V',V} + \langle \partial_t v, C(v)\rangle_{V',V}\Big)\mathrm{d}t.\end{aligned}$$

Using the boundedness of A^{-1}, i.e., $\|A(t)^{-1}\|_{\mathcal{L}(V';V)} \le \alpha^{-1}$ for a.e. $t \in J$ (see Lemma C.51), the boundedness of A_{s}, the integration by parts formula from Lemma 64.40, and the definition of the constant θ, we infer that

$$\begin{aligned}\|\partial_t v + A(v)\|_{L^2(J;V')}^2 \le \frac{1}{\gamma}\Big(&\frac{1}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + M_{\mathrm{s}}\|v\|_{L^2(J;V)}^2 + \|v(T)\|_L^2\\ &+ 2\theta\|\partial_t v\|_{L^2(J;V')}\|v\|_{L^2(J;V)}\Big) - \frac{1}{\gamma}\|v(0)\|_L^2.\end{aligned}$$

Since $\alpha \le M_{\mathrm{s}}$, Young's inequality gives

$$\|\partial_t v + A(v)\|_{L^2(J;V')}^2 \le \frac{1+\theta}{\gamma}\Big(\frac{1}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + M_{\mathrm{s}}\|v\|_{L^2(J;V)}^2 + \|v(T)\|_L^2\Big) - \frac{1}{\gamma}\|v(0)\|_L^2.$$

Since $\gamma\alpha \le \gamma M_{\mathrm{s}} \le 1$, we infer that

$$\|\partial_t v + A(v)\|_{L^2(J;V')}^2 \le \frac{1+\theta}{\gamma^2}\|v\|_X^2 - \frac{1}{\gamma}\|v(0)\|_L^2.$$

Since $\alpha \le \gamma^{-1}$, we have $\|\partial_t v + A(v)\|_{L^2(J;V')}^2 + \alpha\|v(0)\|_L^2 \le \frac{1+\theta}{\gamma^2}\|v\|_X^2$, and this concludes the proof. □

Let us now establish the inf-sup stability of the bilinear form b.

Lemma 71.2 (Inf-sup condition). *Let $\tilde\theta := \frac{1}{2}\operatorname{ess\,sup}_{t\in J}\|\tilde C(t)\|_{\mathcal{L}(V')}$ with $\tilde C(t) := A(t)A(t)^{-*} - I_{V'} \in \mathcal{L}(V')$. The following holds true:*

$$\inf_{v\in X}\sup_{y\in Y}\frac{|b(v,y)|}{\|v\|_X\|y\|_Y} \ge \beta := \alpha\Big(\frac{\alpha\gamma}{1+\tilde\theta}\Big)^{\frac{1}{2}} > 0. \tag{71.8}$$

Proof. Let $v \in X$ and set $y_v := (v(0), A^{-*}(\partial_t v) + v)$. By applying Lemma 65.1 to $J \ni t \mapsto A(t)^{-*} \in \mathcal{L}(V';V)$, we infer that the second component of y_v is strongly measurable, and Bochner's theorem (Theorem 64.12) implies that this component is in $L^2(J;V)$. Moreover, $v(0) \in L$ owing to Lemma 64.40. Hence, $y_v \in Y$, i.e., y_v is an admissible test function. This yields

$$\begin{aligned}
b(v,y_v) &= \int_J \langle \partial_t v + A(v), A^{-*}(\partial_t v) + v\rangle_{V',V}\mathrm{d}t + \|v(0)\|_L^2\\
&= \int_J \Big(\langle A(v), v\rangle_{V',V} + \langle \partial_t v, A^{-1}(\partial_t v)\rangle_{V',V} + 2\langle \partial_t v, v\rangle_{V',V}\Big)\mathrm{d}t + \|v(0)\|_L^2\\
&\ge \alpha\|v\|_{L^2(J;V)}^2 + \gamma\|\partial_t v\|_{L^2(J;V')}^2 + \|v(T)\|_L^2 = \alpha\|v\|_X^2,
\end{aligned}$$

where we used the coercivity of $A(t)$ and $A(t)^{-1}$ for a.e. $t \in J$ and the integration by parts formula from Lemma 64.40. Using the coercivity of $A(t)$ for a.e. $t \in J$, rearranging the terms, and using again the integration by parts formula from Lemma 64.40, we also infer that

$$\begin{aligned}
\|y_v\|_Y^2 &= \|A^{-*}(\partial_t v) + v\|_{L^2(J;V)}^2 + \frac{1}{\alpha}\|v(0)\|_L^2\\
&\le \frac{1}{\alpha}\int_J \langle A(A^{-*}(\partial_t v)) + A(v), A^{-*}(\partial_t v) + v\rangle_{V',V}\mathrm{d}t + \frac{1}{\alpha}\|v(0)\|_L^2\\
&\le \frac{1}{\alpha}\int_J \Big(\langle \partial_t v, A^{-*}(\partial_t v)\rangle_{V',V} + \langle A(v), v\rangle_{V',V}\Big)\mathrm{d}t\\
&\quad + \frac{2\tilde\theta}{\alpha}\|\partial_t v\|_{L^2(J;V')}\|v\|_{L^2(J;V)} + \frac{1}{\alpha}\|v(T)\|_L^2.
\end{aligned}$$

Using Young's inequality to bound the term involving $\tilde\theta$ and the boundedness of $A(t)^{-*}$ and $A_{\mathrm{s}}(t)$ for a.e. $t \in J$, we obtain

$$\|y_v\|_Y^2 \le \frac{1+\tilde\theta}{\alpha}\Big(M_{\mathrm{s}}\|v\|_{L^2(J;V)}^2 + \frac{1}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2\Big) + \frac{1}{\alpha}\|v(T)\|_L^2.$$

Since $\gamma\alpha \le \gamma M_{\mathrm{s}} \le 1$, this yields $\|y_v\|_Y^2 \le \frac{1+\tilde\theta}{\alpha\gamma}\|v\|_X^2$. Putting everything together, we obtain

$$\sup_{y\in Y}\frac{|b(v,y)|}{\|y\|_Y} \ge \frac{|b(v,y_v)|}{\|y_v\|_Y} \ge \beta\,\|v\|_X,$$

with β as defined in the assertion. This concludes the proof. □

Remark 71.3 (Self-adjoint case). If A takes self-adjoint values, we have $\gamma = \frac{1}{M}$ owing to Lemma C.63, so that the X-norm becomes

$$\|v\|_X^2 := \|v\|_{L^2(J;V)}^2 + \frac{1}{\alpha M}\|\partial_t v\|_{L^2(J;V')}^2 + \frac{1}{\alpha}\|v(T)\|_L^2. \tag{71.9}$$

Since $\theta = \tilde\theta = 0$, the boundedness and inf-sup constants of b estimated in (71.7) and (71.8) are $M_b^{\mathrm{s}} := M$ and $\beta^{\mathrm{s}} := \alpha(\frac{\alpha}{M})^{\frac12}$. If $A(t)$ is not self-adjoint, we have $\theta,\tilde\theta \in [0,\frac{M_{\mathrm{ss}}}{\alpha}]$, $M_{\mathrm{ss}} := \operatorname{ess\,sup}_{t\in J}\|A_{\mathrm{ss}}(t)\|_{\mathcal{L}(V;V')}$, $A_{\mathrm{ss}}(t) := \frac12(A(t)-A(t)^*)$, and $\gamma \in [\frac{\alpha}{M^2},\frac{1}{M_{\mathrm{s}}}]$ owing to Lemma C.64, so that $M_b \in [M_{\mathrm{s}},(1+\frac{M_{\mathrm{ss}}}{\alpha})^{\frac12}\frac{M}{\alpha}M]$ and $\beta \in [(1+\frac{M_{\mathrm{ss}}}{\alpha})^{-\frac12}\frac{\alpha}{M}\alpha,(\frac{\alpha}{M_{\mathrm{s}}})^{\frac12}\alpha]$. □

Example 71.4 (Heat equation). Let us consider the heat equation with unit diffusivity ($\kappa := 1$). Then we have $\alpha = \gamma = M = 1$, $\theta = \tilde\theta = 0$, and $\|v\|_X^2 := \|v\|_{L^2(J;H_0^1)}^2 + \|\partial_t v\|_{L^2(J;H^{-1})}^2 + \|v(T)\|_{L^2}^2$, and $b(v,y) := (v(0),y_0)_{L^2} + \int_J\big(\langle\partial_t v(t),y_1(t)\rangle_{H^{-1},H_0^1} + (\nabla v(t),\nabla y_1(t))_{\boldsymbol{L}^2}\big)\mathrm{d}t$. The inf-sup condition (71.8) becomes (see Ern et al. [125] and Exercise 71.3)

$$\|v\|_X^2 \le \sup_{y\in Y}\frac{b(v,y)^2}{\|y_0\|_{L^2}^2 + \|y_1\|_{L^2(J;H_0^1)}^2}, \qquad \forall v\in X.$$ □

71.1.3 Another proof of Lions' theorem

We now reprove Lions' theorem, that is, the parabolic model problem (65.10) is well-posed under the assumption (65.5) (see Theorem 65.9). This is equivalent to saying that the operator $B : X \to Y' = L\times L^2(J;V')$ s.t. $\langle B(v),y\rangle_{Y',Y} := b(v,y)$ is an isomorphism.

Proof. We prove the assertion using the BNB theorem. Since the boundedness and the inf-sup stability of the bilinear form b have already been established in §71.1.2, it only remains to prove that the condition (BNB2) holds true. Let $y := (y_0,y_1) \in Y$ be such that $b(v,y) = 0$ for all $v \in X$. Let $\phi \in C_0^\infty(J;\mathbb{R})$ and $z \in V$. The function $v_0 : J \ni t \mapsto v_0(t) := \phi(t)z \in V$ is in X with $\partial_t v_0(t) = \phi'(t)z$ (see Exercise 71.1). Since the function v_0 vanishes at the initial time, we obtain

$$\begin{aligned}0 = b(v_0,y) &= \int_J\big(\phi'(t)(z,y_1(t))_L + \phi(t)\langle A(t)(z),y_1(t)\rangle_{V',V}\big)\mathrm{d}t\\ &= \int_J \phi(t)\langle -\partial_t y_1(t) + A(t)^*(y_1(t)),z\rangle_{V',V}\mathrm{d}t.\end{aligned}$$

Since ϕ is arbitrary in $C_0^\infty(J;\mathbb{R})$ and z is arbitrary in V, we infer that $\partial_t y_1(t) = A(t)^*(y_1(t))$ for a.e. $t \in J$, which in particular shows that $\partial_t y_1 \in L^2(J;V')$. Let us now use the function $v_1 : J \ni t \mapsto v_1(t) := tz \in V$. Notice that $v_1 \in X$ with $\partial_t v_1(t) = z$, and $v_1(0) = 0$. Integrating by parts in time (Lemma 64.40), we infer that

$$\begin{aligned}0 = b(v_1,y_1) &= \int_J\big((z,y_1(t))_L + t\langle A(t)^*y_1(t),z\rangle_{V',V}\big)\mathrm{d}t\\ &= \int_J\big((z,y_1(t))_L + t\langle\partial_t y_1(t),z\rangle_{V',V}\big)\mathrm{d}t = T(z,y_1(T))_L.\end{aligned}$$

Since z is arbitrary in V and V is dense in L, we obtain $y_1(T) = 0$. We finally use the function $v_2 : J \ni t \mapsto v_2(t) := ty_1$. Notice that $v_2 \in X$ with $\partial_t v_2(t) = \partial_t y_1(t) + t\partial_t y_1(t)$ (see Exercise 71.1) and $v_2(0) = 0$. Since $y_1(T) = 0$, we have

$$\int_J (\partial_t v_2(t), y_1(t))_L \mathrm{d}t = \int_J -\frac{1}{2} t \partial_t \|y_1(t)\|_L^2 \mathrm{d}t = \int_J \frac{1}{2}\|y_1(t)\|_L^2 \mathrm{d}t.$$

Using the coercivity property (65.5c), we infer that

$$0 = b(v_2, y_1) \geq \int_J \Big(\frac{1}{2}\|y_1(t)\|_L^2 + \alpha t \|y_1(t)\|_V^2\Big)\mathrm{d}t,$$

which yields $y_1 = 0$. Therefore, we have $(v(0), y_0)_L = 0$ for all $v \in X$. Considering constant functions in time shows that $(v, y_0)_L = 0$ for all $v \in V$. Since V is dense in L, we have $y_0 = 0$. In conclusion, we have shown that $y = (y_0, y_1) = 0$, i.e., (BNB2) holds true. □

Remark 71.5 (Literature). Theorem 65.9 has been established in Lions [218, Thm. 2.1, p. 219]; see also Lions and Magenes [220, Thm. 4.1, p. 238] or Dautray and Lions [100, Thm. 2, p. 513]. The proof using an inf-sup condition and the BNB theorem has been presented in a previous book by the authors [117], and later in Schwab and Stevenson [250]. Sharp estimates of the inf-sup constant are discussed in Urban and Patera [279], Tantardini and Veeser [270] for parabolic operators and in Ern et al. [125] for the heat equation. □

71.1.4 Ultraweak formulation

Recall from §65.1.5 that in the ultraweak formulation the trial space is $X_{\mathrm{uw}} := L^2(J;V)$ and the test space is $Y_{\mathrm{uw}} := \{w \in L^2(J;V) \mid \partial_t w \in L^2(J;V'),\ w(T) = 0\}$. The ultraweak formulation consists of seeking $u \in X_{\mathrm{uw}}$ s.t. $b_{\mathrm{uw}}(u, w) = \ell_{\mathrm{uw}}(w)$ for all $w \in Y_{\mathrm{uw}}$, with $b_{\mathrm{uw}}(v, w) := \int_J \langle v(t), -\partial_t w(t) + A^*(w)(t)\rangle_{V,V'} \mathrm{d}t$ and $\ell_{\mathrm{uw}}(w) := (u_0, w(0))_L + \int_J \langle f(t), w(t)\rangle_{V',V} \mathrm{d}t$. We equip the trial space X_{uw} with the norm $\|v\|_{X_{\mathrm{uw}}} := \|v\|_{L^2(J;V)}$ and the test space Y_{uw} with the norm (compare with (71.6a))

$$\|w\|_{Y_{\mathrm{uw}}}^2 := \|w\|_{L^2(J;V)}^2 + \frac{\gamma}{\alpha}\|\partial_t w\|_{L^2(J;V')}^2 + \frac{1}{\alpha}\|w(0)\|_L^2. \tag{71.10}$$

Then one can show that (see Exercise 71.4)

$$\inf_{v \in X_{\mathrm{uw}}} \sup_{w \in Y_{\mathrm{uw}}} \frac{|b_{\mathrm{uw}}(v, w)|}{\|v\|_{X_{\mathrm{uw}}}\|w\|_{Y_{\mathrm{uw}}}} \geq \beta > 0, \tag{71.11}$$

where β is the same constant as in the inf-sup condition (71.8). Since Lemma 65.8 asserts that the weak formulation and the ultraweak formulation have the same solution sets, Lions' theorem implies that both formulations are well-posed. In particular, the operator $B_{\mathrm{uw}} : X_{\mathrm{uw}} \to Y_{\mathrm{uw}}'$ s.t. $\langle B_{\mathrm{uw}}(v), w\rangle_{Y_{\mathrm{uw}}', Y_{\mathrm{uw}}} := b_{\mathrm{uw}}(v, w)$ is an isomorphism (see also [125]).

71.2 Semi-discretization in space

We now adopt a point of view slightly more abstract than the pragmatic approach from §66.3 and revisit the error analysis of the semi-discrete problem (66.6), which we recall is formulated as follows:

$$\begin{cases} \text{Find } u_h \in X_h \text{ such that} \\ b(u_h, y_h) = \ell(y_h), \quad \forall y_h \in Y_h. \end{cases} \tag{71.12}$$

Our goal is to derive an error estimate that is quasi-optimal by invoking inf-sup stability. We want to bound the error by the best-approximation error with both the error and the best-approximation error measured in the same norm (see §26.3.2 for the definition of quasi-optimality).

Since the bilinear form b satisfies an inf-sup condition on $X\times Y$ (see Lemma 71.2) and since the approximation setting from §66.2 is conforming (see (66.5)), one may wonder whether b restricted to $X_h\times Y_h$ satisfies an inf-sup condition, uniformly w.r.t. $h\in\mathcal{H}$, when the spaces X_h and Y_h are equipped with the induced norms. We are going to see that the answer to this question is somewhat subtle.

71.2.1 Mesh-dependent inf-sup stability

Since $V_h\subset V$, we equip V_h with the V-norm, and since V_h is finite-dimensional, we identify $V_h\equiv V_h'$ by means of the inner product in L. Let $A_h(t):V_h\to V_h$ be the discrete operator s.t. $(A_h(t)(v_h),w_h)_L:=a(t;v_h,w_h)$ for all $v_h,w_h\in V_h$ and a.e. $t\in J$. Then $A_h(t)$ is coercive and bounded, uniformly w.r.t. $h\in\mathcal{H}$ and a.e. $t\in J$, with the constants α and M, respectively. To reproduce the arguments from the proof of Lemma 71.2, one needs to invoke the coercivity of $A_h(t)^{-1}$ on V_h. For this property to hold true uniformly w.r.t. $h\in\mathcal{H}$, we consider the following additional norm on V_h:

$$\|\phi_h\|_{V_h'}:=\sup_{v_h\in V_h}\frac{|(\phi_h,v_h)_L|}{\|v_h\|_V},\qquad\forall\phi_h\in V_h. \tag{71.13}$$

Let us set

$$\gamma_h:=\operatorname*{ess\,inf}_{t\in J}\inf_{\phi_h\in V_h}\frac{(\phi_h,A_h(t)^{-1}(\phi_h))_L}{\|\phi_h\|_{V_h'}^2}. \tag{71.14}$$

The argument to prove (C.29) shows that $\gamma_h\ge\frac{\alpha_h}{M_h^2}$, where α_h and M_h are, respectively, coercivity and boundedness constants of A_h that are uniform w.r.t. $t\in J$. Since $\alpha\le\alpha_h\le M_h\le M$, we infer that $\gamma_h\ge\frac{\alpha}{M^2}>0$, i.e., γ_h is bounded from below away from zero uniformly w.r.t. $h\in\mathcal{H}$.

Lemma 71.6 (Mesh-dependent inf-sup condition). *Let us equip Y_h with the Y-norm defined in* (71.6b) *and $X_h:=H^1(J;V_h)$ with*

$$\|v_h\|_{X_h}^2:=\|v_h\|_{L^2(J;V)}^2+\frac{\gamma_h}{\alpha}\|\partial_t v_h\|_{L^2(J;V_h')}^2+\frac{1}{\alpha}\|v_h(T)\|_L^2. \tag{71.15}$$

Let $\tilde\theta_h:=\frac12\operatorname{ess\,sup}_{t\in J}\|A_h(t)A_h(t)^{-*}-I_{V_h'}\|_{\mathcal{L}(V_h')}$ and $\beta_h:=\alpha(\frac{\alpha\gamma_h}{1+\tilde\theta_h})^{\frac12}$, then

$$\inf_{v_h\in X_h}\sup_{y_h\in Y_h}\frac{|b(v_h,y_h)|}{\|v_h\|_{X_h}\|y_h\|_Y}\ge\beta_h>0. \tag{71.16}$$

Proof. Proceed as in the proof of Lemma 71.2 and use that $\alpha\le\alpha_h$. □

Remark 71.7 (Value of β_h). Proceeding as in Remark 71.3, we infer that $\beta_h\ge\beta_\flat:=(1+\frac{M_{\mathrm{ss}}}{\alpha})^{-\frac12}\frac{\alpha}{M}\alpha$, i.e., β_h is bounded from below away from zero uniformly w.r.t. $h\in\mathcal{H}$. If A, and hence A_h, take selfadjoint values, Lemma C.63 implies that $\gamma_h=\frac{1}{M_h}$ and $\gamma=\frac{1}{M}$, so that $\gamma_h\ge\gamma$. Since $\tilde\theta_h=0$ in this case, we conclude that $\beta_h\ge\alpha(\frac{\alpha}{M})^{\frac12}=\beta^{\mathrm{S}}$, where β^{S} is the inf-sup constant of b on $X\times Y$ in the self-adjoint case. □

71.2.2 Inf-sup stability in the X-norm

The discrete inf-sup condition from Lemma 71.6 is not entirely satisfactory since the mesh-dependent norm $\|\cdot\|_{V_h'}$ is used to measure the time derivative. To avoid this situation, one needs to equip the subspace X_h with the norm of X. The key question is then to compare the norms $\|\cdot\|_{V_h'}$ and $\|\cdot\|_{V'}$ on V_h.

Lemma 71.8 (Comparison of dual norms). *The following holds true:*

$$0 < \frac{1}{\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}} = \inf_{\phi_h \in V_h} \frac{\|\phi_h\|_{V_h'}}{\|\phi_h\|_{V'}} \le \sup_{\phi_h \in V_h} \frac{\|\phi_h\|_{V_h'}}{\|\phi_h\|_{V'}} \le 1. \tag{71.17}$$

Proof. The value for the infimum is the identity (26.23) in Example 26.22 (see Lemma 26.19 and Tantardini and Veeser [270, Thm. 2.1]):

$$\frac{1}{\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}} = \inf_{\phi_h \in V_h} \sup_{w_h \in V_h} \frac{(\phi_h, w_h)_L}{\|\phi_h\|_{V'}\|w_h\|_V} = \inf_{\phi_h \in V_h} \frac{\|\phi_h\|_{V_h'}}{\|\phi_h\|_{V'}}.$$

The upper bound on the supremum is obtained by extending the supremizing set from V_h to V in the definition (71.13). □

Lemma 71.8 means that the uniform V-stability of the L-orthogonal projection is a necessary and sufficient condition for the uniform equivalence of the norms $\|\cdot\|_{V_h'}$ and $\|\cdot\|_{V'}$ on V_h (see also Andreev [10, Lem. 6.2]). In the context of the heat equation where $V := H_0^1(D)$ and $L := L^2(D)$, sufficient conditions on the mesh sequence for the uniform H^1-stability of the L^2-orthogonal projection are identified in Remark 22.23. In the rest of this section, we assume that there is $c_{\mathcal{P}}$ s.t. for all $h \in \mathcal{H}$,

$$\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)} \le c_{\mathcal{P}}. \tag{71.18}$$

Lemma 71.9 (Inf-sup condition with X-norm). *Let X_h be equipped with the X-norm and let Y_h be equipped with the Y-norm. The following holds true with $\beta_h' := \beta_h \|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}^{-1} (\frac{\gamma_h}{\gamma})^{\frac{1}{2}}$ and β_h defined in Lemma* 71.6:

$$\inf_{v_h \in X_h} \sup_{y_h \in Y_h} \frac{|b(v_h, y_h)|}{\|v_h\|_X \|y_h\|_Y} \ge \beta_h' > 0. \tag{71.19}$$

Proof. The lower bound in (71.17) implies that

$$\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}^2 \frac{\gamma}{\gamma_h} \frac{\gamma_h}{\alpha} \|\partial_t v_h\|_{L^2(J;V_h')}^2 \ge \frac{\gamma}{\alpha} \|\partial_t v_h\|_{L^2(J;V')}^2.$$

Hence, $\|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)} (\frac{\gamma}{\gamma_h})^{\frac{1}{2}} \|v_h\|_{X_h} \ge \|v_h\|_X$. Recalling the definition of the X-norm in (71.6a) and the definition of the X_h-norm in (71.15), we infer that (71.16) implies (71.19). □

Remark 71.10 (Value of β_h'). Under the assumption (71.18), β_h' is bounded from below away from zero uniformly w.r.t. $h \in \mathcal{H}$, since we have $\beta_h' \ge \beta_{\flat} c_{\mathcal{P}}^{-1} (\frac{\alpha}{M} \frac{M_{\mathrm{s}}}{M})^{\frac{1}{2}}$, where we used that $\frac{\alpha}{M^2} \le \gamma_h$ and $\gamma \le \frac{1}{M_{\mathrm{s}}}$, and where $\beta_{\flat} > 0$ is defined in Remark 71.7. In the self-adjoint case, we have $\gamma_h \ge \gamma$ and $\beta_h \ge \alpha (\frac{\alpha}{M})^{\frac{1}{2}} = \beta^{\mathrm{s}}$, so that $\beta_h' \ge \beta^{\mathrm{s}} c_{\mathcal{P}}^{-1}$. □

Theorem 71.11 (Quasi-optimal error estimate, X-norm). *Let $u \in X$ solve* (71.2) *and let $u_h \in X_h$ solve* (71.12)*. The following quasi-optimal error bound holds true:*

$$\|u - u_h\|_X \le \left(1 + \frac{M_b}{\beta_h'}\right) \inf_{v_h \in X_h} \|u - v_h\|_X. \tag{71.20}$$

Proof. Combine the abstract error estimate from Lemma 26.14 (see also Lemma 26.18) with the inf-sup stability from Lemma 71.9 and the boundedness of b on $X\times Y$ (recall that the approximation setting is conforming). □

In the case of the heat equation, where A takes self-adjoint values, we have $\gamma=\frac{1}{M}$, $M_b=M$ (see Lemma 71.1), and $\beta^{\mathrm{S}}=\alpha(\frac{\alpha}{M})^{\frac12}$ (see Lemma 71.2) with $\alpha:=\kappa_\flat$, $M:=\kappa_\sharp$, and $0<\kappa_\flat\le\kappa\le\kappa_\sharp$ in $D\times J$. Note that $\frac{M_b}{\beta'_h}\le(\frac{M}{\alpha})^{\frac32}c_{\mathcal P}$. As mentioned in Remark 71.3, the X-norm is $\|v\|_X^2:=\|v\|_{L^2(J;H^1_0)}^2+\frac{1}{\alpha M}\|\partial_t v\|_{L^2(J;H^{-1})}^2+\frac1\alpha\|v(T)\|_{L^2}^2$. Recall the time scale $\rho:=\frac{2}{C_{\mathrm{PS}}^2}\frac{\ell_D^2}{\alpha}$.

Corollary 71.12 (Convergence rate, heat equation). *Let $r\in[1,k]$, where $k\ge1$ is the polynomial degree of the finite elements used to build the discrete space V_h. Assume that $u\in L^2(J;H^{r+1}(D))\cap H^1(J;H^{r-1}(D))$. Under the assumption* (71.18) *and letting $\xi_\kappa:=\frac{M}{\alpha}$ and $\xi_\rho:=\frac{\rho}{T}$, the following holds true for all $h\in\mathcal H$,*

$$\|u-u_h\|_X\le c\,\xi_\kappa^{\frac32}h^r\biggl(\int_J\Bigl(\max(\xi_\rho,\xi_\kappa^{\frac12})|u(t)|_{H^{r+1}(D)}^2+\xi_\kappa^{\frac12}\frac{1}{\alpha M}|\partial_t u(t)|_{H^{r-1}(D)}^2\Bigr)\mathrm{d}t\biggr)^{\frac12}.\tag{71.21}$$

Proof. We use $v_h(t):=\mathcal P_{V_h}(u(t))$ in the right-hand side of (71.20) for a.e. $t\in J$. Thus, we need to bound $\|\eta\|_X$ with $\eta(t):=u(t)-\mathcal P_{V_h}(u(t))$. Letting $\phi(t):=\frac{t}{T}$, we have

$$\frac12\|\eta(T)\|_{L^2}^2=\int_J\langle\partial_t(\phi\eta),\phi\eta\rangle_{H^{-1},H^1_0}\mathrm{d}t=\int_J\bigl(\phi'\phi\|\eta\|_{L^2}^2+\phi^2\langle\partial_t\eta,\eta\rangle_{H^{-1},H^1_0}\bigr)\mathrm{d}t.$$

Using Young's inequality and $\xi_\kappa^{-1}=\frac{\alpha}{M}$, we obtain

$$|\langle\partial_t\eta,\eta\rangle_{H^{-1},H^1_0}|\le\|\eta\|_{H^1_0}\|\partial_t\eta\|_{H^{-1}}\le\frac{\alpha}{2}\xi_\kappa^{\frac12}\Bigl(\|\eta\|_{H^1_0}^2+\frac{1}{\alpha M}\|\partial_t\eta\|_{H^{-1}}^2\Bigr).$$

Since $T\|\phi'\|_{L^\infty(J)}=\|\phi\|_{L^\infty(J)}=1$, $C_{\mathrm{PS}}\|\eta\|_{L^2}\le\ell_D\|\eta\|_{H^1_0}$ owing to the Poincaré–Steklov inequality, and $\xi_\rho=\frac{\rho}{T}$, we infer that $\int_J\phi'\phi\|\eta\|_{L^2}^2\mathrm{d}t\le\frac{\alpha}{2}\frac{\rho}{T}\|\eta\|_{L^2(H^1_0)}^2$. Putting the above bounds together leads to

$$\frac1\alpha\|\eta(T)\|_{L^2}^2\le(\xi_\rho+\xi_\kappa^{\frac12})\|\eta\|_{L^2(J;H^1_0)}^2+\xi_\kappa^{\frac12}\frac{1}{\alpha M}\|\partial_t\eta\|_{L^2(J;H^{-1})}^2.$$

Thus, we have proved that

$$\|\eta\|_X^2\le(1+\xi_\rho+\xi_\kappa^{\frac12})\|\eta\|_{L^2(J;H^1_0)}^2+(1+\xi_\kappa^{\frac12})\frac{1}{\alpha M}\|\partial_t\eta\|_{L^2(J;H^{-1})}^2,$$

and it remains to bound the two terms on the right-hand side. We invoke Proposition 22.21 for the first term. For the second term, we observe that $\partial_t\mathcal P_{V_h}(u)=\mathcal P_{V_h}(\partial_t u)$ and that for all $\eta\in L^2(D)$, $\|\eta-\mathcal P_{V_h}(\eta)\|_{H^{-1}(D)}\le ch\|\eta-\mathcal P_{V_h}(\eta)\|_{L^2(D)}$ (see Exercise 22.6). Combined with Proposition 22.19 this implies that $\|\eta-\mathcal P_{V_h}(\eta)\|_{H^{-1}(D)}\le ch^r|\eta|_{H^{r-1}(D)}$ for all $r\in[1,k]$. This proves (71.21) since $1\le\xi_\kappa$ and $1+\frac{M_b}{\beta'_h}\le c\xi_\kappa^{\frac32}$. □

Remark 71.13 (Localization in space). The upper bound in (71.21) is not localized over the mesh cells because we used the L^2-orthogonal projection to bound $\inf_{v_h\in X_h}\|\partial_t(u-v_h)\|_{L^2(J;H^{-1}(D))}$. Using a variant of the Scott–Zhang interpolation operator that preserves mean-values over element patches, it is possible to localize this upper bound over the mesh cells. We refer the reader to Remark 66.15 and [270, p. 337] for more details. As done in §66.3, it is also possible to localize the upper bound by making the slightly stronger smoothness assumption $u\in L^2(J;H^{r+1}(D))\cap H^1(J;H^r(D))$ and by using the bound $\|\partial_t\eta\|_{L^2(J;H^{-1}(D))}\le C_{\mathrm{PS}}^{-1}\ell_D\|\partial_t\eta\|_{L^2(J;L^2(D))}$, which is a consequence of the Poincaré–Steklov inequality. □

Remark 71.14 (Literature). Quasi-optimal error estimates using the discrete norm $\|\cdot\|_{X_h}$ have been derived by Dupont [113]. Quasi-optimal error estimates in the X-norm have been established by Chrysafinos and Hou [86] under the assumption (71.18) requiring that the L-orthogonal projection be uniformly V-stable. That this assumption is not only sufficient but also necessary for quasi-optimality in the X-norm has been proved in [270]. □

Remark 71.15 ($C^0(\overline{J};L^2(D))$-estimate). The estimate (71.21) gives the suboptimal convergence order $\mathcal{O}(h^r)$ on $\|u-u_h\|_{C^0(\overline{J};L^2(D))}$ under the smoothness assumption $u\in L^2(J;H^{r+1}(D))$ and $\partial_t u\in L^2(J;H^{r-1}(D))$. Using the elliptic projection and under the same smoothness assumption on u but with $\partial_t u\in L^2(J;H^{r+1}(D))$ and other appropriate assumptions if the bilinear form a is time-dependent, the estimate (66.18) gives $\mathcal{O}(h^{r+s})$ where s is the index of elliptic regularity pickup. This extra smoothness requirement appears to be the price to pay to achieve optimal error decay rates. It is also possible to combine the use of the elliptic projection with inf-sup stability to obtain an error estimate in the $C^0(\overline{J};L)$-norm; see Exercise 71.7. □

A quasi-optimal error bound in the $L^2(J;V)$-norm on the solution to the semi-discrete problem (71.12) can be established by invoking the ultraweak formulation (see §71.1.4).

Theorem 71.16 (Quasi-optimal $L^2(J;V)$-error estimate). *Let $u\in X$ solve* (71.2) *and let $u_h\in X_h$ solve* (71.12). (i) *Under the assumption* (71.18) *there is c s.t. for all $h\in\mathcal{H}$,*

$$\|u-u_h\|_{L^2(J;V)}\le c\inf_{v_h\in L^2(J;V_h)}\|u-v_h\|_{L^2(J;V)}. \tag{71.22}$$

(ii) *Assuming for the heat equation that $u\in L^2(J;H^{r+1}(D))$, we have*

$$\|u-u_h\|_{L^2(J;H_0^1(D))}\le c\left(\int_J\sum_{K\in\mathcal{T}_h}h_K^{2r}|u(t)|^2_{H^{r+1}(K)}\mathrm{d}t\right)^{\frac12}. \tag{71.23}$$

Proof. (i) Recall from Lemma 65.8 that the solution to (65.10) is in $X_{\mathrm{uw}}:=L^2(J;V)$ and satisfies $b_{\mathrm{uw}}(u,w)=\ell_{\mathrm{uw}}(w)$ for all $w\in Y_{\mathrm{uw}}:=\{w\in L^2(J;V)\mid \partial_t w\in L^2(J;V'),\ w(T)=0\}$. Reasoning as in the proof of Lemma 65.8, one can show that the semi-discrete solution u_h to the problem (66.6) is in $X_{\mathrm{uw},h}:=L^2(J;V_h)$ and satisfies $b_{\mathrm{uw}}(u_h,w_h)=\ell_{\mathrm{uw}}(w_h)$ for all $w_h\in Y_{\mathrm{uw},h}:=\{w_h\in H^1(J;V_h)\mid w_h(T)=0\}$. The formulation (66.6) is a conforming approximation of the ultraweak formulation since $X_{\mathrm{uw},h}\subset X_{\mathrm{uw}}$ and $Y_{\mathrm{uw},h}\subset Y_{\mathrm{uw}}$. By proceeding as in §71.1.4 and Exercise 71.4, we deduce that the inf-sup condition (71.19) implies that

$$\inf_{v_h\in X_{\mathrm{uw},h}}\sup_{y_h\in Y_{\mathrm{uw},h}}\frac{|b_{\mathrm{uw}}(v_h,y_h)|}{\|v_h\|_{X_{\mathrm{uw}}}\|y_h\|_{Y_{\mathrm{uw}}}}\ge\beta_h'>0,$$

with $\|v\|_{X_{\mathrm{uw}}}:=\|v\|_{L^2(J;V)}$ and $\|w\|^2_{Y_{\mathrm{uw}}}:=\|w\|^2_{L^2(J;V)}+\frac{\gamma}{\alpha}\|\partial_t w\|^2_{L^2(J;V')}+\frac{1}{\alpha}\|w(0)\|^2_L$. Since b_{uw} is bounded on $X_{\mathrm{uw}}\times Y_{\mathrm{uw}}$ using the above norms, we can now invoke the abstract error estimate from Lemma 26.14 (see also Lemma 26.18) to infer that (71.22) holds true.
(ii) (71.23) readily follows from (71.22). □

Remark 71.17 (Estimate (71.23)**).** The advantage of the estimate (71.23) w.r.t. (71.21) is that (71.23) only requires optimal smoothness on u, but does not assume anything on $\partial_t u$; see also [270, p. 338]. (Notice though that, letting κ be constant for simplicity, the identity $\partial_t u-f=\kappa\Delta u$ implies that assuming $u\in L^2(J;H^r(D))$ is equivalent to assuming $\partial_t u-f\in L^2(J;H^{r-2}(D))$.) Obviously, (71.23) is less informative than (71.21) since the latter also bounds the error on the time derivative and estimates the error pointwise in time. □

71.3 dG(k) scheme

Let $k \ge 0$. In this section, we revisit the dG(k) scheme introduced in Chapter 69 and strengthen the stability analysis for this scheme by proving an inf-sup condition. Our goal is to obtain a discrete counterpart of (71.2) for the bilinear form b_τ defined in (69.13).

To avoid distracting technicalities, we assume that the bilinear form a is symmetric. We identify $V_h \equiv V_h'$ by means of the L-inner product, and we define $A_h(t) : V_h \to V_h$ s.t. $(A_h(t)(v_h), w_h)_L = a(t; v_h, w_h)$ for all $v_h, w_h \in V_h$ and all $t \in (0, T]$. Recalling that $\mathcal{R}_\tau$ is the reconstruction operator defined in (69.17), we equip the spaces $X_{h\tau} := P_k^{\mathrm{b}}(\overline{J}_\tau; V_h)$ and $Y_{h\tau} := X_{h\tau}$ (see (69.12)) with the following norms:

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \|v_{h\tau}\|_{L^2(J;V)}^2 + \frac{1}{\alpha M}\|\partial_t \mathcal{R}_\tau(v_{h\tau})\|_{L^2(J;V_h')}^2 + \frac{1}{\alpha}\|v_{h\tau}(T)\|_L^2 + \frac{1}{\alpha}\sum_{n\in\mathcal{N}_\tau}\|[\![v_{h\tau}]\!]_{n-1}\|_L^2,$$

$$\|y_{h\tau}\|_{Y_{h\tau}}^2 := \|v_{h\tau}\|_{L^2(J;V)}^2 + \frac{1}{\alpha}\|y_{h\tau}(0)\|_L^2,$$

where $\alpha > 0$ and $M < \infty$ are the coercivity and boundedness parameters of the bilinear form a. Recall that $\|v_h\|_{V_h'} := \sup_{w_h\in V_h} \frac{|(v_h,w_h)_L|}{\|w_h\|_V}$. Notice that $\|\cdot\|_{X_{h\tau}}$ defines a norm on $X_{h\tau}$. Indeed, $\|v_{h\tau}\|_{X_{h\tau}} = 0$ implies that $v_{h\tau|J_\tau} = 0$, and $v_{h\tau}(0) = 0$ follows from $[\![v_{h\tau}]\!]_0 = 0$. (Notice that the coercivity norm considered in Lemma 69.15 is slightly different.)

Lemma 71.18 (Inf-sup stability). *Assume that the bilinear form a is symmetric. The following holds true:*

$$\inf_{v_{h\tau}\in X_{h\tau}} \sup_{y_{h\tau}\in Y_{h\tau}} \frac{|b_\tau(v_{h\tau}, w_{h\tau})|}{\|v_{h\tau}\|_{X_{h\tau}}\|y_{h\tau}\|_{Y_{h\tau}}} \ge \alpha\Big(\frac{\alpha}{M}\Big)^{\frac{1}{2}}. \tag{71.24}$$

Proof. (1) Let $v_{h\tau} \in X_{h\tau}$ and let us set $r_{h\tau} := \mathcal{R}_\tau(v_{h\tau})$ for conciseness. Since the function $r_{h\tau}$ is globally continuous in time and piecewise smooth, we have $\partial_t r_{h\tau} \in L^1(J; V_h)$. Owing to (69.18) and the identity derived in the proof of Lemma 69.15, we also infer that

$$\begin{aligned}\int_J (\partial_t r_{h\tau}, v_{h\tau})\mathrm{d}t &= \sum_{n\in\mathcal{N}_\tau}\Big(\int_{J_n}(\partial_t v_{h\tau}, v_{h\tau})_L\mathrm{d}t + ([\![v_{h\tau}]\!]_{n-1}, v_{h\tau}(t_{n-1}^+))_L\Big)\\ &= \frac{1}{2}\|v_{h\tau}(T)\|_L^2 - \frac{1}{2}\|v_{h\tau}(0)\|_L^2 + \sum_{n\in\mathcal{N}_\tau}\frac{1}{2}\|[\![v_{h\tau}]\!]_{n-1}\|_L^2,\end{aligned}$$

where we dropped the time dependency to simplify the notation.
(2) Since we assumed that a is symmetric, i.e., $A_h^{-1} = A_h^{-*}$, we would like to consider the test function $A_h^{-1}(\partial_t r_{h\tau}) + v_{h\tau}$, but unfortunately this is not a polynomial function in time if the operator A_h is time-dependent. To fix this issue, we invoke the interpolation operator $\mathcal{I}_k^{\mathrm{GR}}$ defined in (69.9). Thus, we consider the partner $y_{h\tau} \in Y_{h\tau}$ s.t. $y_{h\tau}(0) := v_{h\tau}(0)$ and $y_{h\tau}(t) := \mathcal{I}_k^{\mathrm{GR}}(A_h^{-1}(\partial_t r_{h\tau}))(t) + v_{h\tau}(t)$ for all $t \in J_\tau$. Recalling the identity (69.18) we have $b_\tau(v_{h\tau}, y_{h\tau}) =: \mathfrak{T}_1 + \mathfrak{T}_2 + \|v_{h\tau}(0)\|_L^2$, where

$$\begin{aligned}\mathfrak{T}_1 &:= \int_J (\partial_t r_{h\tau}, \mathcal{I}_k^{\mathrm{GR}}(A_h^{-1}(\partial_t r_{h\tau})) + v_{h\tau})_L\mathrm{d}t,\\ \mathfrak{T}_2 &:= \int_J (A_h(v_{h\tau}), \mathcal{I}_k^{\mathrm{GR}}(A_h^{-1}(\partial_t r_{h\tau})) + v_{h\tau})_L\,\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t).\end{aligned}$$

We have $\mathfrak{T}_1 = \int_J(\partial_t r_{h\tau}, A_h^{-1}(\partial_t r_{h\tau}))_L\,\mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) + \int_J(\partial_t r_{h\tau}, v_{h\tau})_L\mathrm{d}t$ owing to (69.11b). Since $A_h^{-1}(t_{n,l})$ is γ_h-coercive on V_h' with $\gamma_h \ge \gamma = \frac{1}{M}$ (see Remark 71.7) and since the weights of

the quadrature are all positive, we infer that

$$\int_{J_n} (\partial_t r_{h\tau}, A_h^{-1}(\partial_t r_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \geq \sum_{l\in\{1:k+1\}} \frac{1}{M}\omega_l \|\partial_t r_{h\tau}(t_{n,l})\|_{V_h'}^2 \geq \frac{1}{M}\|\partial_t r_{h\tau}\|_{L^2(J_n;V_h')}^2,$$

where we used that $\|\partial_t r_{h\tau}\|_{V_h'}^2 \in \mathbb{P}_{2k}(J_n;\mathbb{R})$. This implies that

$$\mathfrak{T}_1 \geq \frac{1}{M}\|\partial_t r_{h\tau}\|_{L^2(J_n;V_h')}^2 + \int_J (\partial_t r_{h\tau}, v_{h\tau})_L \mathrm{d}t.$$

Similarly, owing to (69.11a) we infer that

$$\begin{aligned}
\mathfrak{T}_2 &= \int_J (A_h(v_{h\tau}), A_h^{-1}(\partial_t r_{h\tau}) + v_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\
&= \int_J (v_{h\tau}, \partial_t r_{h\tau} + A_h(v_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\
&= \int_J (v_{h\tau}, \partial_t r_{h\tau})_L \mathrm{d}t + \int_J (v_{h\tau}, A_h(v_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t),
\end{aligned}$$

where we used that $(v_{h\tau}, \partial_t r_{h\tau})_L \in \mathbb{P}_{2k}(J_n;\mathbb{R})$ for all $n \in \mathcal{N}_\tau$. Invoking the α_h-coercivity of $A_h(t_{n,l})$ on V_h with $\alpha_h \geq \alpha$, using the positivity of the quadrature weights, and that $\|v_{h\tau}\|_V^2 \in \mathbb{P}_{2k}(J_n;\mathbb{R})$, we infer that

$$\mathfrak{T}_2 \geq \int_J (\partial_t r_{h\tau}, v_{h\tau})_L \mathrm{d}t + \alpha\|v_{h\tau}\|_{L^2(J;V)}^2.$$

Putting everything together, and recalling the identity from Step (1), yields $b_\tau(v_{h\tau}, y_{h\tau}) \geq \alpha\|v_{h\tau}\|_{X_{h\tau}}^2$.
(3) Using the coercivity of A_h at $t_{n,l}$ for every integers $n \in \mathcal{N}_\tau$ and $l \in \{1{:}k{+}1\}$, we infer that $\alpha\|y_{h\tau}\|_{Y_{h\tau}}^2 \leq \|v_{h\tau}(0)\|_L^2 + \mathfrak{T}_3$ with

$$\begin{aligned}
\mathfrak{T}_3 &:= \int_J (A_h(y_{h\tau}), y_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\
&= \int_J (A_h(A_h^{-1}(\partial_t r_{h\tau}) + v_{h\tau}), A_h^{-1}(\partial_t r_{h\tau}) + v_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t),
\end{aligned}$$

where we used (69.11a). Rearranging the terms and since $(\partial_t r_{h\tau}, v_{h\tau})_L \in \mathbb{P}_{2k}(J_n;\mathbb{R})$ for all $n \in \mathcal{N}_\tau$, we obtain

$$\begin{aligned}
\mathfrak{T}_3 = &\int_J 2(\partial_t r_{h\tau}, v_{h\tau})_L \mathrm{d}t + \int_J (A_h(v_{h\tau}), v_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\
&+ \int_J (A_h^{-1}(\partial_t r_{h\tau}), \partial_t r_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t).
\end{aligned}$$

Recalling that $\|A_h(t_{n,l})(w_h)\|_{V_h'} \leq M\|w_h\|_V$ and $\|A_h(t_{n,l})^{-1}(w_h)\|_{V_h'} \leq \frac{1}{\alpha}\|w_h\|_{V_h'}$ for all $w_h \in V_h$, all $n \in \mathcal{N}_\tau$, and all $l \in \{1{:}k{+}1\}$, and using the identity from Step (1) proves that $\|y_{h\tau}\|_{Y_{h\tau}}^2 \leq \frac{M}{\alpha}\|v_{h\tau}\|_{X_{h\tau}}^2$. Combining this bound with the lower bound from Step (2) gives (71.24). □

Remark 71.19 (Inf-sup condition). The inf-sup condition (71.24) is the counterpart of the condition established for the continuous parabolic problem in Lemma 71.2 with the same constant. The only difference is that the time derivative is now measured using the $\|\cdot\|_{V_h'}$-norm. One can replace this norm by the $\|\cdot\|_{V'}$-norm whenever the L-orthogonal projection onto V_h is uniformly V-stable, as done in §71.2.2. (The uniform stability holds true if the mesh sequence is quasi-uniform, see also Remark 22.23.) The reader is referred to Smears [262], Neumüller and Smears [229] for further results on the inf-sup stability of dG(k) schemes with a time-independent bilinear form a. □

71.4 cPG(k) scheme

Let $k \geq 1$. In this section, we revisit the cPG(k) scheme introduced in Chapter 70 and strengthen the stability analysis for this scheme by proving an inf-sup condition. To do so, we equip the spaces $X_{h\tau} := P_k^{\mathrm{g}}(\overline{J}_\tau; V_h)$ and $Y_{h\tau} := P_{k-1}^{\mathrm{b}}(\overline{J}_\tau; V_h)$ (see (70.6)) with the following norms:

$$\|v_{h\tau}\|_{X_{h\tau}}^2 := \|\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})\|_{L^2(J;V)}^2 + \frac{1}{\alpha M}\|\partial_t v_{h\tau}\|_{L^2(J;V_h')}^2 + \frac{1}{\alpha}\|v_{h\tau}(T)\|_L^2,$$
$$\|y_{h\tau}\|_{Y_{h\tau}}^2 := \|y_{h\tau}\|_{L^2(J;V)}^2 + \frac{1}{\alpha}\|y_{h\tau}(0)\|_L^2,$$

where $\alpha > 0$ and $M < \infty$ are the coercivity and boundedness constants of the bilinear form a. Recall that $\|v_h\|_{V_h'} := \sup_{w_h \in V_h} \frac{|(v_h, w_h)_L|}{\|w_h\|_V}$.

Lemma 71.20 (Inf-sup stability). *The following holds true:*

$$\inf_{v_{h\tau} \in X_{h\tau}} \sup_{y_{h\tau} \in Y_{h\tau}} \frac{|b_\tau(v_{h\tau}, w_{h\tau})|}{\|v_{h\tau}\|_{X_{h\tau}} \|y_{h\tau}\|_{Y_{h\tau}}} \geq \alpha\Big(\frac{\alpha}{M}\Big)^{\frac{1}{2}}. \tag{71.25}$$

Proof. We are going to use the integral identities (70.27). Let $v_{h\tau} \in X_{h\tau}$. Taking inspiration from the proof of Lemma 71.18, we consider the partner $y_{h\tau} \in Y_{h\tau}$ s.t. $y_{h\tau}(0) := v_{h\tau}(0)$ and $y_{h\tau}(t) := \mathcal{I}_{k-1}^{\mathrm{GL}}(A_h^{-1}(\partial_t v_{h\tau}) + v_{h\tau})(t)$ for all $t \in J_\tau$. Notice that indeed $y_{h\tau} \in Y_{h\tau}$. Moreover, we have $b_\tau(v_{h\tau}, y_{h\tau}) =: \mathfrak{T}_1 + \mathfrak{T}_2 + \|v_{h\tau}(0)\|_L^2$ with

$$\begin{aligned}\mathfrak{T}_1 &:= \int_J (\partial_t v_{h\tau}, y_{h\tau})_L \mathrm{d}t = \int_J (\partial_t v_{h\tau}, \mathcal{I}_{k-1}^{\mathrm{GL}}(A_h^{-1}(\partial_t v_{h\tau}) + v_{h\tau}))_L \mathrm{d}t \\ &= \int_J (\partial_t v_{h\tau}, A_h^{-1}(\partial_t v_{h\tau}))_L \, \mu_k^{\mathrm{GL}}(\mathrm{d}t) + \int_J (\partial_t v_{h\tau}, v_{h\tau})_L \mathrm{d}t,\end{aligned}$$

where we used (70.27a) and that $(\partial_t v_{h\tau}, v_{h\tau})_L \in \mathbb{P}_{2k-1}(J_n; \mathbb{R})$ for all $n \in \mathcal{N}_\tau$, and the quadrature is of order $(2k-1)$. Since $A_h^{-1}(t_{n,l})$ is γ_h-coercive on V_h' with $\gamma_h \geq \gamma = \frac{1}{M}$ (see Remark 71.7) and since the weights of the quadrature are all positive, we infer that

$$\int_{J_n} (\partial_t v_{h\tau}, A_h^{-1}(\partial_t v_{h\tau}))_L \, \mu_k^{\mathrm{GL}}(\mathrm{d}t) \geq \sum_{l \in \{1:k\}} \frac{1}{M}\omega_l \|\partial_t v_{h\tau}(t_{n,l})\|_{V_h'}^2 \geq \frac{1}{M}\|\partial_t v_{h\tau}\|_{L^2(J_n;V_h')}^2,$$

where we used that $\|\partial_t v_{h\tau}\|_{V_h'}^2 \in \mathbb{P}_{2k-2}(J_n; \mathbb{R})$. This implies that

$$\mathfrak{T}_1 \geq \frac{1}{M}\|\partial_t v_{h\tau}\|_{L^2(J_n;V_h')}^2 + \int_J (\partial_t v_{h\tau}, v_{h\tau})_L \mathrm{d}t.$$

Similarly, owing to (70.27b) we infer that

$$\begin{aligned}\mathfrak{T}_2 &:= \int_J (A_h(v_{h\tau}), \mathcal{I}_{k-1}^{\mathrm{GL}}(A_h^{-1}(\partial_t v_{h\tau}) + v_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\ &= \int_J (A_h(v_{h\tau}), A_h^{-1}(\partial_t v_{h\tau}) + v_{h\tau})_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\ &= \int_J (v_{h\tau}, \partial_t v_{h\tau} + A_h(v_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t) \\ &= \int_J (v_{h\tau}, \partial_t v_{h\tau})_L \mathrm{d}t + \int_J (v_{h\tau}, A_h(v_{h\tau}))_L \, \mu_{k+1}^{\mathrm{GR}}(\mathrm{d}t),\end{aligned}$$

where we used that $(v_{h\tau}, \partial_t v_{h\tau})_L \in \mathbb{P}_{2k-1}(J_n;\mathbb{R})$ for all $n \in \mathcal{N}_\tau$. Reasoning as in the proof of Lemma 70.8 for the second term on the right-hand side, we infer that

$$\mathfrak{T}_2 \geq \int_J (\partial_t v_{h\tau}, y_{h\tau})_L \mathrm{d}t + \alpha \|\mathcal{I}_{k-1}^{\mathrm{GL}}(v_{h\tau})\|_{L^2(J;V)}^2.$$

Putting everything together and since $2\int_J (\partial_t v_{h\tau}, y_{h\tau})_L \mathrm{d}t = \|v_{h\tau}(T)\|_L^2 - \|v_{h\tau}(0)\|_L^2$ proves that $b_\tau(v_{h\tau}, y_{h\tau}) \geq \alpha \|v_{h\tau}\|_{X_{h\tau}}^2$. Finally, we have $\|y_{h\tau}\|_{Y_{h\tau}} \leq \left(\frac{M}{\alpha}\right)^{\frac12} \|v_{h\tau}\|_{X_{h\tau}}$ (see Exercise 71.9), and combining these two bounds proves the assertion. □

Exercises

Exercise 71.1 (Time derivative). Let $\phi \in C_0^\infty(J;\mathbb{R})$ and $v \in X$, i.e., $v \in L^2(J;V)$ and $\partial_t v \in L^2(J;V')$. Show that ϕv is in X with $\partial_t(\phi v)(t) = \phi'(t)v(t) + \phi(t)\partial_t v(t)$. (*Hint*: use Pettis theorem and Lemma 64.33.)

Exercise 71.2 (Inf-sup condition). Prove (71.7) with X equipped with the norm $\|v\|_X^2 := \|v\|_{L^2(J;V)}^2 + \frac{\gamma}{\alpha}\|\partial_t v\|_{L^2(J;V')}^2 + \gamma\|v(0)\|_L^2$. (*Hint*: use integration by parts in time to bound $\gamma\|v(0)\|_L^2$ by $\|v\|_X^2$.)

Exercise 71.3 (Heat equation). Consider the heat equation with unit diffusivity (see Example 71.4). Prove that for all $v \in X$,

$$\|v\|_X^2 = \sup_{y_1 \in L^2(J;H_0^1)} \frac{b(v,(0,y_1))^2}{\|y_1\|_{L^2(J;H_0^1)}^2} + \|v(0)\|_{L^2}^2.$$

(*Hint*: observe that the supremum is reached for $y_1 := A^{-1}(\partial_t v) + v$.)

Exercise 71.4 (Ultraweak formulation). Equip the space X_{uw} with the norm $\|v\|_{X_{\mathrm{uw}}} := \|v\|_{L^2(J;V)}$ and the space Y_{uw} with the norm defined in (71.10). (i) Prove the inf-sup condition (71.11). (*Hint*: consider the adjoint parabolic problem $\partial w_v(t) + A^*(w_v)(t) := (v(t),\cdot)_V$ for a.e. $t \in J$, with $w_v(0) := 0$, invoke Lemma 71.2, then set $\tilde{w}_v(t) := w_v(T-t)$.) (ii) The rest of the exercise considers the heat equation with unit diffusivity. Show that $\sup_{w\in Y_{\mathrm{uw}}} \frac{b_{\mathrm{uw}}(v,w)}{\|w\|_{Y_{\mathrm{uw}}}} \leq \|v\|_{X_{\mathrm{uw}}}$ for all $v \in X_{\mathrm{uw}}$. (*Hint*: prove first that $\|A^{-1}(\partial_t w) - w\|_{L^2(J;H_0^1(D))}^2 = \|w\|_{Y_{\mathrm{uw}}}^2$ for all $w \in Y_{\mathrm{uw}}$.) (iii) Prove that

$$\|v\|_{X_{\mathrm{uw}}} = \sup_{w\in Y_{\mathrm{uw}}} \frac{b_{\mathrm{uw}}(v,w)}{\|w\|_{Y_{\mathrm{uw}}}}, \qquad \forall v \in X_{\mathrm{uw}}.$$

(*Hint*: compute $b(v,\widetilde{w})$, where $\widetilde{w}_v \in Y_{\mathrm{uw}}$ solve the backward-in-time parabolic problem $-\partial_t \widetilde{w}_v - \Delta \widetilde{w}_v = -\Delta v$ with $\widetilde{w}_v(T) = 0$.)

Exercise 71.5 (Norm $\|\cdot\|_{V_h'}$). Let $\|\cdot\|_{V_h'}$ be defined in (71.13). Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h and let $\mathcal{S} \in \mathbb{R}^{I\times I}$ and $\mathcal{M} \in \mathbb{R}^{I\times I}$ be the stiffness and mass matrices s.t. $\mathcal{S}_{ij} := (\varphi_j,\varphi_i)_V$ and $\mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L$ for all $i,j \in \{1:I\}$ (these matrices are symmetric positive definite). For all $v_h \in V_h$, let $\mathsf{V} \in \mathbb{R}^I$ be the coordinate vector of v_h in the basis $\{\varphi_i\}_{i\in\{1:I\}}$, i.e., $v_h := \sum_{i\in\{1:I\}} \mathsf{V}_i\varphi_i$. (i) Prove that $\|v_h\|_{V_h'} = (\mathsf{V}^\mathsf{T}\mathcal{M}\mathcal{S}^{-1}\mathcal{M}\mathsf{V})^{\frac12}$. (*Hint*: use that $\|v_h\|_{V_h'} = \sup_{\mathsf{W}\in\mathbb{R}^I} \frac{\mathsf{W}^\mathsf{T}\mathcal{M}\mathsf{V}}{(\mathsf{W}^\mathsf{T}\mathcal{S}\mathsf{W})^{\frac12}}$.) (ii) Let $\mu \geq 0$. Prove the following two-sided bound due to Pearson and Wathen [235] (see also Smears [262]):

$$\frac12 \leq \frac{\mathsf{V}^\mathsf{T}(\mathcal{M}\mathcal{S}^{-1}\mathcal{M} + \mu\mathcal{S})\mathsf{V}}{\mathsf{V}^\mathsf{T}(\mathcal{M} + \mu^{\frac12}\mathcal{S})\mathcal{S}^{-1}(\mathcal{M} + \mu^{\frac12}\mathcal{S})\mathsf{V}} \leq 2, \qquad \forall \mathsf{V} \in \mathbb{R}^I.$$

Exercise 71.6 (Error analysis with $\|\cdot\|_{X_h}$). Referring to §71.2 and denoting by u_h the solution to (71.12), let $\eta(t) := u(t) - \mathcal{P}_{V_h}(u(t))$ for a.e. $t \in J$. (i) With the norm $\|\cdot\|_{X_h}$ defined in (71.15), prove that $|b(\eta, y_h)| \le \sqrt{2}M\|\eta\|_{X_h}\|y_h\|_Y$ for all $y_h \in Y_h$. (*Hint*: use that $\frac{\alpha}{\gamma_h} \le M^2$.) (ii) Prove the error estimate $\|u - u_h\|_{X_h} \le \Big(1 + \frac{\sqrt{2}M}{\beta_h}\Big)\|\eta\|_{X_h}$, where β_h is the constant from the inf-sup inequality (71.16). (*Hint*: combine inf-sup stability with consistency and boundedness.)

Exercise 71.7 ($C^0(\overline{J};L)$-estimate using inf-sup stability). (i) Recalling that $\|\cdot\|_X$ is defined in (71.6a), prove that $\gamma^{\frac12}\|v\|_{C^0(\overline{J};L^2)} \le \|v\|_X$ for all $v \in X$. (*Hint*: see Exercise 71.2.) (ii) Assume (71.18). Let $c_1 := \sqrt{\frac{\gamma}{\alpha}}$ and $c_2 := \sqrt{\frac{\rho}{2}}$, where $\rho := 2\frac{\iota_{L,V}^2}{\alpha}$ and $\iota_{L,V}$ is the operator norm of the embedding $V \hookrightarrow L$, i.e., the smallest constant s.t. $\|v\|_L \le \iota_{L,V}\|v\|_V$ for all $v \in V$. Prove that

$$\beta_h' c_1\|u - u_h\|_{C^0(\overline{J};L)} \le \beta_h' c_1\|\eta\|_{C^0(\overline{J};L)} + \|\eta(0)\|_L + c_2\|\partial_t\eta\|_{L^2(J;L)},$$

with $\eta(t) := u(t) - \Pi_h^{\mathrm{E}}(t;u(t))$. (*Hint*: combine Lemma 71.9 with consistency.) (iii) Compare this estimate with (66.16) in the context of the heat equation.

Exercise 71.8 (Implicit Euler scheme). (i) Let $X_{h\tau} := (V_h)^{N+1}$ and $Y_{h\tau} := V_h \times (V_h)^N$. Reformulate the implicit Euler scheme (67.3) using the forms

$$\begin{aligned} b_\tau(v_{h\tau}, y_{h\tau}) &:= (v_h^0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\big(((\delta_\tau v_{h\tau})^n, y_{1h}^n)_L + a^n(v_h^n, y_{1h}^n)\big),\\ \ell_\tau(y_{h\tau}) &:= (u_0, y_{0h})_L + \sum_{n\in\mathcal{N}_\tau} \tau\langle f^n, y_{1h}^n\rangle_{V',V}, \end{aligned}$$

where $(\delta_\tau v_{h\tau})^n := \frac{1}{\tau}(v_h^n - v_h^{n-1})$. (ii) Assume that the bilinear form a is symmetric at all times. Prove that

$$\alpha\|u_{h\tau}\|_{\ell^2(J;V)}^2 + \frac{1}{M}\|\delta_\tau u_{h\tau}\|_{\ell^2(J;V_h')}^2 + \tau\|\delta_\tau u_{h\tau}\|_{\ell^2(J;L)}^2 + \|u_h^N\|_L^2 \le \frac{M}{\alpha}\Big(\frac{1}{\alpha}\|f\|_{\ell^2(J;V')}^2 + \|u_0\|_L^2\Big).$$

(*Hint*: use the inf-sup condition (67.27).) (iii) Assume that $u \in C^0(\overline{J};V) \cap C^1(\overline{J};V') \cap H^2(J;V')$ and that $\mathcal{P}_{V_h}$ is uniformly V-stable (see (71.18)). Prove that

$$\begin{aligned} \|\delta_\tau u_{h\tau} - \delta_\tau u_\tau\|_{\ell^2(J;V')} \le \|\mathcal{P}_{V_h}\|_{\mathcal{L}(V)}\frac{M}{\alpha}\Bigg(&\sqrt{3}\big(M\|\eta_\tau\|_{\ell^2(J;V)} + 2\|\partial_t\eta\|_{L^2(J;V')}\\ &+ \tau\|\partial_{tt}u\|_{L^2(J;V')}\big) + \sqrt{\alpha}\|e_h^0\|_L\Bigg), \end{aligned}$$

where $(\delta_\tau u_\tau)^n := \frac{1}{\tau}(u(t_n) - u(t_{n-1}))$ for all $n \in \mathcal{N}_\tau$, $\eta(t) := u(t) - v_h(t)$ for all $t \in \overline{J}$, $\eta_\tau := (\eta(t_n))_{n\in\mathcal{N}_\tau}$, and v_h arbitrary in $H^1(J;V_h)$. (*Hint*: use Step (ii) and Lemma 71.8.)

Exercise 71.9 (Inf-sup for cPG(k)). Complete the proof of Lemma 71.20. (*Hint*: reason as in the last step of the proof of Lemma 71.18.)

Chapter 72

Weak formulations and well-posedness

The four chapters composing Part XIV deal with the time-dependent version of the steady Stokes problem investigated in Chapter 53. The present chapter focuses on the weak formulation of the time-dependent Stokes equations. We consider two possible weak formulations. The first one enforces the divergence-free constraint on the velocity field without introducing the pressure. This formulation can be handled by using the same analysis tools as for parabolic problems. The second weak formulation includes the pressure. This formulation entails some subtleties concerning the smoothness in time of the pressure and of the time derivative of the velocity. Both formulations hinge on the Bochner integration theory exposed in Chapter 64. The next three chapters deal with the approximation of the mixed weak formulation in space and in time. The discretization in space relies on stable mixed finite elements, and the approximation in time relies on either monolithic or fractional-step schemes.

72.1 Model problem

Let $J := (0,T)$, $T > 0$, be the time interval and D be a Lipschitz domain in $\mathbb{R}^d$. We want to model the time-dependent flow of an incompressible fluid in D assuming that the inertial forces are negligible. Let $\boldsymbol{f} : D\times J \to \mathbb{R}^d$ be a vector-valued field (the body force acting on the fluid) and $\boldsymbol{u}_0$ be a divergence-free velocity field (the initial velocity field). Let $\partial D = \partial D_{\mathrm{d}} \cup \partial D_{\mathrm{n}}$ be a partition of the boundary, and assume for simplicity that $|\partial D_{\mathrm{d}}| > 0$. The time-dependent Stokes problem consists of seeking the velocity field $\boldsymbol{u} : D\times J \to \mathbb{R}^d$ and the pressure field $p : D\times J \to \mathbb{R}$ such that

$$\partial_t \boldsymbol{u} - \nabla\cdot\mathbb{s}(\boldsymbol{u}) + \nabla p = \boldsymbol{f} \qquad \text{in } D\times J, \tag{72.1a}$$

$$\nabla\cdot\boldsymbol{u} = 0 \qquad \text{in } D\times J, \tag{72.1b}$$

$$\boldsymbol{u}_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}, \quad (\mathbb{s}(\boldsymbol{u})\boldsymbol{n} - p\boldsymbol{n})_{|\partial D_{\mathrm{n}}} = \boldsymbol{a}_{\mathrm{n}} \qquad \text{on } \partial D\times J, \tag{72.1c}$$

$$\boldsymbol{u}(\cdot,0) = \boldsymbol{u}_0(\cdot) \qquad \text{in } D. \tag{72.1d}$$

The equations (72.1a)-(72.1b) express the balance of momentum and mass, respectively (note that the mass balance does not involve any time derivative owing to the incompressible nature of the motion). The equation (72.1c) enforces the boundary conditions, and (72.1d) enforces the initial

condition (the velocity field is prescribed initially, but the pressure is not). The second-order tensor $\mathbb{s}(\boldsymbol{u})$ in (72.1a) is the viscous stress tensor defined as

$$\mathbb{s}(\boldsymbol{u}) := 2\mu\mathbb{e}(\boldsymbol{u}), \qquad \mathbb{e}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathsf{T}}), \tag{72.2}$$

where $\mathbb{e}(\boldsymbol{u})$ is the (linearized) strain rate tensor and $\mu > 0$ is the dynamic viscosity. For simplicity, we assume that μ is constant. In the steady-state situation, the time derivative $\partial_t\boldsymbol{u}$ vanishes in (72.1a) and the initial condition (72.1d) becomes irrelevant, i.e., we recover the steady Stokes equations studied in Chapters 53–55.

Remark 72.1 (Dirichlet condition, bulk viscosity). As for the steady Stokes equations, one can consider the non-homogeneous Dirichlet condition $\boldsymbol{u}_{|\partial D_{\mathrm{d}}} = \boldsymbol{a}_{\mathrm{d}}$. One recovers the homogeneous condition by introducing a suitable lifting of $\boldsymbol{a}_{\mathrm{d}}$ (see Remark 53.6). One can also define the viscous stress tensor as $\mathbb{s}'(\boldsymbol{u}) := 2\mu\mathbb{e}(\boldsymbol{u}) + \lambda(\nabla\cdot\boldsymbol{u})\mathbb{I}$, where $\lambda \geq 0$ is the bulk viscosity and $\mathbb{I}$ the $d\times d$ identity tensor. Then (72.1a) becomes $\partial_t\boldsymbol{u} - \nabla\cdot\mathbb{s}'(\boldsymbol{u}) + \nabla p' = \boldsymbol{f}$ with the pressure redefined as $p' := p + \lambda\nabla\cdot\boldsymbol{u}$ (see Remark 53.5). □

Let us briefly recall the functional setting for the steady Stokes equations (see §53.2.1). The trial and test space for the velocity is the Hilbert space

$$\boldsymbol{V}_{\mathrm{d}} := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \boldsymbol{v}_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}\}, \tag{72.3}$$

where $\boldsymbol{v}_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}$ means more precisely that $\gamma^{\mathrm{g}}(\boldsymbol{v})_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}$ and $\gamma^{\mathrm{g}} : \boldsymbol{H}^1(D) \to \boldsymbol{H}^{\frac{1}{2}}(\partial D)$ is the trace map acting componentwise. Since $|\partial D_{\mathrm{d}}| > 0$, Korn's second inequality (see (42.14)) implies that there is C_{K} such that $C_{\mathrm{K}}\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)} \leq \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}$. The Poincaré–Steklov inequality (applied componentwise) then implies that there is $C_{\mathrm{KPS}} > 0$ such that

$$C_{\mathrm{KPS}}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \leq \ell_D\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}, \qquad \forall\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}. \tag{72.4}$$

We equip the velocity space $\boldsymbol{V}_{\mathrm{d}}$ with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$ (we could also use the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}} := \|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)} = |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ as in the steady case). Moreover, recalling that the pressure is defined up to an additive constant if only Dirichlet conditions are prescribed, i.e., when $\partial D_{\mathrm{d}} = \partial D$ (see Remark 53.4), the trial and test space for the pressure is the Hilbert space

$$Q := \begin{cases} L^2(D) & \text{if } \partial D \neq \partial D_{\mathrm{d}}, \\ L^2_*(D) := \{q \in L^2(D) \mid \int_D q\,\mathrm{d}x = 0\} & \text{if } \partial D = \partial D_{\mathrm{d}}. \end{cases} \tag{72.5}$$

We equip the space Q with the L^2-norm and we identify Q with Q'. We define the bounded bilinear forms $a : \boldsymbol{V}_{\mathrm{d}}\times\boldsymbol{V}_{\mathrm{d}} \to \mathbb{R}$ and $b : \boldsymbol{V}_{\mathrm{d}}\times Q \to \mathbb{R}$ such that

$$a(\boldsymbol{v},\boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}):\mathbb{e}(\boldsymbol{w})\,\mathrm{d}x, \qquad b(\boldsymbol{v},q) := -\int_D q\nabla\cdot\boldsymbol{v}\,\mathrm{d}x. \tag{72.6}$$

These bilinear forms satisfy the following coercivity property and the following inf-sup condition (see Lemma 53.9):

$$\inf_{\boldsymbol{v}\in\boldsymbol{V}_{\mathrm{d}}} \frac{|a(\boldsymbol{v},\boldsymbol{v})|}{\|\boldsymbol{v}\|^2_{\boldsymbol{V}_{\mathrm{d}}}} \geq \alpha := 2\mu, \qquad \inf_{q\in Q}\sup_{\boldsymbol{v}\in\boldsymbol{V}_{\mathrm{d}}} \frac{|b(\boldsymbol{v},q)|}{\|q\|_Q\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}}} \geq \beta > 0. \tag{72.7}$$

Recall that the inf-sup condition means that the operator $B : \boldsymbol{V}_{\mathrm{d}} \to Q$ s.t. $(B(\boldsymbol{v}),q)_{L^2(D)} := b(\boldsymbol{v},q) = -(\nabla\cdot\boldsymbol{v},q)_{L^2(D)}$ is surjective.

72.2 Constrained weak formulation

To account for the incompressibility constraint, we consider the subspaces

$$\boldsymbol{\mathcal{H}}_{\rm d} := \{\boldsymbol{v} \in \boldsymbol{L}^2(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D_{\rm d}}\cdot\boldsymbol{n} = 0\}, \tag{72.8a}$$

$$\boldsymbol{\mathcal{V}}_{\rm d} := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D_{\rm d}} = \boldsymbol{0}\}, \tag{72.8b}$$

where in (72.8a) $\nabla\cdot\boldsymbol{v} = 0$ and $\boldsymbol{v}_{|\partial D_{\rm d}}\cdot\boldsymbol{n} = 0$ mean that $(\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(D)} = 0$ for all $q \in H^1(D)$ s.t. $\gamma^{\rm g}(q)_{|\partial D_{\rm n}} = 0$ (or for all $q \in H^1(D)$ s.t. $(q,1)_{L^2(D)} = 0$ if $\partial D = \partial D_{\rm d}$). Notice that $\boldsymbol{\mathcal{V}}_{\rm d} = \{\boldsymbol{v} \in \boldsymbol{V}_{\rm d} \mid \nabla\cdot\boldsymbol{v} = 0\} = \ker(B)$. One important property of the pair $(\boldsymbol{\mathcal{V}}_{\rm d}, \boldsymbol{\mathcal{H}}_{\rm d})$ is that the space $\boldsymbol{\mathcal{V}}_{\rm d}$ is dense and embeds continuously in $\boldsymbol{\mathcal{H}}_{\rm d}$. Thus, we have a Gelfand triple $\boldsymbol{\mathcal{V}}_{\rm d} \hookrightarrow \boldsymbol{\mathcal{H}}_{\rm d} \equiv \boldsymbol{\mathcal{H}}_{\rm d}' \hookrightarrow \boldsymbol{\mathcal{V}}_{\rm d}'$. These statements are established in Temam [272, pp. 15-18] if $\partial D_{\rm d} = \partial D$.

The constrained weak formulation of the time-dependent Stokes equations hinges on the above Gelfand triple and the abstract setting for parabolic problems introduced in §65.1.2, where $\boldsymbol{\mathcal{V}}_{\rm d}$ plays the role of V and $\boldsymbol{\mathcal{H}}_{\rm d}$ plays the role of the pivot space L. Thus, we consider the functional space (see (64.6))

$$X(J; \boldsymbol{\mathcal{V}}_{\rm d}, \boldsymbol{\mathcal{V}}_{\rm d}') := \{\boldsymbol{v} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}) \mid \partial_t\boldsymbol{v} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}')\}. \tag{72.9}$$

Let us assume that the data satisfies $\boldsymbol{f} \in L^2(J; \boldsymbol{V}_{\rm d}')$ and $\boldsymbol{a}_{\rm n} \in L^2(J; \boldsymbol{L}^2(\partial D_{\rm n}))$. Then by setting

$$\langle \boldsymbol{F}(t), \boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}}_{\rm d}', \boldsymbol{\mathcal{V}}_{\rm d}} := \langle \boldsymbol{f}(t), \boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}}_{\rm d}', \boldsymbol{\mathcal{V}}_{\rm d}} + \int_{\partial D_{\rm n}} \boldsymbol{a}_{\rm n}(t)\cdot\boldsymbol{w}\,{\rm d}s, \tag{72.10}$$

for a.e. $t \in J$ and all $\boldsymbol{w} \in \boldsymbol{\mathcal{V}}_{\rm d}$, we define a linear form $\boldsymbol{F} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}') \equiv L^2(J; \boldsymbol{\mathcal{V}}_{\rm d})'$ (see Theorem 64.20(i)). Notice that the action of $\boldsymbol{f}(t)$ on $\boldsymbol{w}$ is meaningful since $\boldsymbol{f}(t) \in \boldsymbol{V}_{\rm d}'$ by assumption and $\boldsymbol{w} \in \boldsymbol{\mathcal{V}}_{\rm d} \subset \boldsymbol{V}_{\rm d}$. Moreover, assuming $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}}_{\rm d}$, we infer from Lemma 64.40(i) that the initial condition $\boldsymbol{u}(0) := \boldsymbol{u}_0$ is meaningful whenever $\boldsymbol{u} \in X(J; \boldsymbol{\mathcal{V}}_{\rm d}, \boldsymbol{\mathcal{V}}_{\rm d}')$. The constrained weak formulation is as follows:

$$\left\{\begin{array}{l} \text{Find } \boldsymbol{u} \in X(J; \boldsymbol{\mathcal{V}}_{\rm d}, \boldsymbol{\mathcal{V}}_{\rm d}') \text{ s.t. } \boldsymbol{u}(0) = \boldsymbol{u}_0 \text{ and for all } \boldsymbol{w} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}), \\ \int_J \big(\langle \partial_t\boldsymbol{u}, \boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}}_{\rm d}', \boldsymbol{\mathcal{V}}_{\rm d}} + a(\boldsymbol{u}, \boldsymbol{w})\big){\rm d}t = \int_J \langle \boldsymbol{F}, \boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}}_{\rm d}', \boldsymbol{\mathcal{V}}_{\rm d}}{\rm d}t. \end{array}\right. \tag{72.11}$$

The initial condition can also be enforced by means of a test function $\boldsymbol{y}_0 \in \boldsymbol{\mathcal{H}}_{\rm d}$ as we did for parabolic problems.

Proposition 72.2 (Well-posedness). *Assume that $\boldsymbol{F} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}')$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}}_{\rm d}$. Then the model problem* (72.11) *is well-posed.*

Proof. We apply the well-posedness theory for parabolic problems (see Theorem 65.9). We consider the Gelfand triple $(\boldsymbol{\mathcal{V}}_{\rm d}, \boldsymbol{\mathcal{H}}_{\rm d} \equiv \boldsymbol{\mathcal{H}}_{\rm d}', \boldsymbol{\mathcal{V}}_{\rm d}')$. The operator $\widetilde{A} : \boldsymbol{\mathcal{V}}_{\rm d} \to \boldsymbol{\mathcal{V}}_{\rm d}'$ associated with the bilinear form $\widetilde{a} := a_{|\boldsymbol{\mathcal{V}}_{\rm d}\times\boldsymbol{\mathcal{V}}_{\rm d}}$ satisfies the hypotheses (65.5). Moreover, $\boldsymbol{F} \in L^2(J; \boldsymbol{\mathcal{V}}_{\rm d}')$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}}_{\rm d}$ by assumption. □

72.3 Mixed weak formulation with smooth data

The well-posedness statement in Proposition 72.2 is somewhat unsatisfactory since it does not give any information on the pressure. The objective of this section is to fill this gap. To simplify the argumentation, we assume from now on that $\partial D_{\rm d} = \partial D$, and we simplify the notation by using

$\boldsymbol{V} := \boldsymbol{H}^1_0(D)$ instead of $\boldsymbol{V}_{\rm d}$ and $\boldsymbol{\mathcal{V}} := \{\boldsymbol{v} \in \boldsymbol{V} \mid \nabla{\cdot}\boldsymbol{v} = 0\}$ instead of $\boldsymbol{\mathcal{V}}_{\rm d}$. Both spaces are equipped with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathfrak{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$.

In the previous section, we assumed that $\boldsymbol{f} \in L^2(J;\boldsymbol{H}^{-1}(D))$ and

$$\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}} := \{\boldsymbol{v} \in \boldsymbol{L}^2(D) \mid \nabla{\cdot}\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D}{\cdot}\boldsymbol{n} = 0\},$$

and we established the existence and uniqueness of the weak solution in $X(J;\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{V}}')$, i.e., $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$ and $\partial_t\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}}')$. In the present section, we assume a bit more smoothness on the data, i.e., $\boldsymbol{f} \in L^2(J;\boldsymbol{L}^2(D))$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{V}}$ (a setting with less regularity is treated in §72.4). We consider the following mixed weak formulation: Find $\boldsymbol{u} \in X(J;\boldsymbol{H}^1_0(D),\boldsymbol{L}^2(D))$ and $p \in L^2(J;L^2_*(D))$ s.t. $\boldsymbol{u}(0) := \boldsymbol{u}_0$ and

$$\begin{cases} (\partial_t\boldsymbol{u}(t),\boldsymbol{w})_{\boldsymbol{L}^2} + a(\boldsymbol{u}(t),\boldsymbol{w}) + b(\boldsymbol{w},p(t)) = (\boldsymbol{f}(t),\boldsymbol{w})_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}(t),q) = 0, \end{cases} \tag{72.12}$$

for all $\boldsymbol{w} \in \boldsymbol{H}^1_0(D)$ and all $q \in L^2_*(D)$, where the two equalities hold in $L^2(J)$. Notice that in (72.12) we have $\partial_t\boldsymbol{u} \in L^2(J;\boldsymbol{L}^2(D))$. Moreover, the second equation in (72.12) implies that $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$. An equivalent restatement of (72.12) is that for all $\boldsymbol{y} \in L^2(J;\boldsymbol{H}^1_0(D))$ and all $r \in L^2(J;L^2_*(D))$,

$$\int_J \big((\partial_t\boldsymbol{u},\boldsymbol{y})_{\boldsymbol{L}^2} + a(\boldsymbol{u},\boldsymbol{y}) + b(\boldsymbol{y},p) - b(\boldsymbol{u},r)\big)\mathrm{d}t = \int_J (\boldsymbol{f},\boldsymbol{y})_{\boldsymbol{L}^2}\mathrm{d}t. \tag{72.13}$$

To prove the well-posedness of (72.12), we are going to use the well-posedness of the constrained weak formulation (72.11), establish some a priori estimates on the time derivative of the velocity, and deduce the existence of the pressure in $L^2(J;L^2_*(D))$. Recall that the bilinear forms a and b satisfy the coercivity property and the inf-sup condition stated in (72.7). To simplify the notation, we introduce the time scale $\rho := C_{\rm KPS}^{-2}\frac{\ell_D^2}{\mu}$.

Theorem 72.3 (Well-posedness, a priori estimates). *Assume $\boldsymbol{f} \in L^2(J;\boldsymbol{L}^2(D))$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{V}}$.*
(i) *The mixed weak formulation* (72.12) *is well-posed, and the following a priori estimates hold true:*

$$2\mu\|\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{V})} \le \tfrac12\rho\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + \|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}, \tag{72.14a}$$

$$\|\partial_t\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu\|\boldsymbol{u}_0\|^2_{\boldsymbol{V}}, \tag{72.14b}$$

$$\|p\|^2_{L^2(J;L^2)} \le \frac{1}{\beta^2}\rho\mu\big(10\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 12\mu\|\boldsymbol{u}_0\|^2_{\boldsymbol{V}}\big). \tag{72.14c}$$

(ii) *We have for all $t \in (0,T]$ with $J_t := (0,t)$,*

$$\|\boldsymbol{u}(t)\|^2_{\boldsymbol{L}^2} \le \tfrac12\rho\|e^{-\frac{t-\cdot}{\rho}}\boldsymbol{f}\|^2_{L^2(J_t;\boldsymbol{L}^2)} + e^{-2\frac{t}{\rho}}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}. \tag{72.15}$$

Proof. (1) Estimates on $\boldsymbol{u}$. Owing to Proposition 72.2, there is $\boldsymbol{u} \in X(J;\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{V}}')$ solving the constrained weak formulation (72.11) with the right-hand side replaced by $(\boldsymbol{f}(t),\boldsymbol{v})_{\boldsymbol{L}^2}$. The estimate (72.14a) is obtained by proceeding as in Lemma 65.10: one uses the test function $\boldsymbol{w} := \boldsymbol{u}$ in (72.11) and invokes the coercivity of a, Korn's inequality, and Young's inequality. Moreover, the estimate (72.15) is obtained by using the test function $\boldsymbol{w}(t) := e^{2\frac{t}{\rho}}\boldsymbol{u}(t)$ in (72.11) and by proceeding as in Lemma 65.11.
(2) Estimate on $\partial_t\boldsymbol{u}$. We proceed as in the proof of Lemma 65.13. Let $(\boldsymbol{v}_i)_{i\in\mathbb{N}}$ be a Hilbert basis of $\boldsymbol{\mathcal{V}}$ (recall that $\boldsymbol{\mathcal{V}} \subset \boldsymbol{H}^1_0(D)$). Let $n \in \mathbb{N}$ and set $\boldsymbol{\mathcal{V}}_n := \mathrm{span}\{\boldsymbol{v}_i\}_{i\in\{0:n\}}$. Let $\boldsymbol{u}_{0n}$ be the $\boldsymbol{V}$-orthogonal projection of $\boldsymbol{u}_0$ onto $\boldsymbol{\mathcal{V}}_n$. We consider the following set of ordinary differential equations:

$$(\partial_t\boldsymbol{u}_n(t),\boldsymbol{v})_{\boldsymbol{L}^2} + a(\boldsymbol{u}_n(t),\boldsymbol{v}) = (\boldsymbol{f}(t),\boldsymbol{v})_{\boldsymbol{L}^2}, \tag{72.16}$$

for all $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_n$ and a.e. $t \in J$, supplemented with the initial condition $\boldsymbol{u}_n(0) = \boldsymbol{u}_{0n}$. Owing to the Cauchy–Lipschitz theorem, (72.16) has a unique solution $\boldsymbol{u}_n \in X(J; \boldsymbol{\mathcal{V}}_n, \boldsymbol{\mathcal{V}}_n)$. Moreover, using the test function $\boldsymbol{v}_n := \partial_t \boldsymbol{u}_n(t)$ in (72.16), for a.e. $t \in J$, and integrating over the time interval J leads to

$$\|\partial_t \boldsymbol{u}_n\|^2_{L^2(J;\boldsymbol{L}^2)} + \frac{1}{2} a(\boldsymbol{u}_n(T), \boldsymbol{u}_n(T)) = \int_J (\boldsymbol{f}(t), \partial_t \boldsymbol{u}_n(t))_{\boldsymbol{L}^2} \mathrm{d}t + \frac{1}{2} a(\boldsymbol{u}_{0n}, \boldsymbol{u}_{0n}).$$

Invoking the coercivity and the boundedness of a, the Cauchy–Schwarz and Young's inequalities, and the bound $\|\boldsymbol{u}_{0n}\|_{\boldsymbol{V}} \le \|\boldsymbol{u}_0\|_{\boldsymbol{V}}$, we infer that

$$\|\partial_t \boldsymbol{u}_n\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu \|\boldsymbol{u}_n(T)\|^2_{\boldsymbol{V}} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu \|\boldsymbol{u}_0\|^2_{\boldsymbol{V}}. \tag{72.17}$$

Moreover, integrating from 0 to t for all $t \in (0, T]$ shows that $2\mu \|\boldsymbol{u}\|^2_{L^\infty(J;\boldsymbol{V})} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu \|\boldsymbol{u}_0\|^2_{\boldsymbol{V}}$. The estimate (72.17), which crucially hinges on the assumptions on $\boldsymbol{f}$ and $\boldsymbol{u}_0$, shows that the sequence $(\partial_t \boldsymbol{u}_n)_{n\in\mathbb{N}}$ is bounded in $L^2(J; \boldsymbol{L}^2(D))$. Similarly, testing (72.16) with $\boldsymbol{u}_n$ and proceeding as in Lemma 65.10 shows that the sequence $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ is bounded in $L^2(J; \boldsymbol{\mathcal{V}})$. Hence, there is a subsequence (that we do not renumber for simplicity) such that $\partial_t \boldsymbol{u}_n \rightharpoonup \boldsymbol{w}_*$ weakly in $L^2(J; \boldsymbol{L}^2(D))$ and $\boldsymbol{u}_n \rightharpoonup \boldsymbol{u}_*$ weakly in $L^2(J; \boldsymbol{\mathcal{V}})$ as $n \to \infty$. Uniqueness of the weak limit implies that $\partial_t \boldsymbol{u}_* = \boldsymbol{w}_*$, showing that $\boldsymbol{u}_* \in X(J; \boldsymbol{\mathcal{V}}, \boldsymbol{L}^2(D))$. Moreover, fixing $m \in \mathbb{N}$ and a test function $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_m$, we can pass to the limit $n \to \infty$ in (72.16) and show that $(\partial_t \boldsymbol{u}_*(t), \boldsymbol{v})_{\boldsymbol{L}^2} + a(\boldsymbol{u}_*(t), \boldsymbol{v}) = (\boldsymbol{f}(t), \boldsymbol{v})_{\boldsymbol{L}^2}$ in $L^2(J)$. Since $\boldsymbol{v}$ is arbitrary in $\boldsymbol{\mathcal{V}}_m$, m is arbitrary in $\mathbb{N}$, and the family $\{\boldsymbol{\mathcal{V}}_m\}_{m\in\mathbb{N}}$ is dense in $\boldsymbol{\mathcal{V}}$, the above equality holds for every test function $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$. Furthermore, (72.17) shows that the sequence $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ is bounded in $X^{\infty,2}(J; \boldsymbol{\mathcal{V}}, \boldsymbol{L}^2(D)) := \{\boldsymbol{v} \in L^\infty(J; \boldsymbol{\mathcal{V}}) \mid \partial_t \boldsymbol{v} \in L^2(J; \boldsymbol{L}^2(D))\}$. The compactness result from Theorem 64.39(ii) then implies that, up to a subsequence, $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ converges in $C^0(\overline{J}; \boldsymbol{L}^2(D))$. By uniqueness of the limit, we infer that $\boldsymbol{u}_n(0) \to \boldsymbol{u}_*(0)$ in $\boldsymbol{L}^2(D)$ as $n \to \infty$, and since $\boldsymbol{u}_{0n} \to \boldsymbol{u}_0$, we conclude that $\boldsymbol{u}_*(0) = \boldsymbol{u}(0)$. Invoking the uniqueness of the solution to the constrained weak formulation (72.11) shows that $\boldsymbol{u}_* = \boldsymbol{u}$. This proves that $\boldsymbol{u} \in X(J; \boldsymbol{\mathcal{V}}, \boldsymbol{L}^2(D)) \subset X(J; \boldsymbol{H}^1_0(D), \boldsymbol{L}^2(D))$ and that $\|\partial_t \boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)} \le \limsup_{n\to\infty} \|\partial_t \boldsymbol{u}_n\|^2_{L^2(J;\boldsymbol{L}^2)} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu \|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}$, i.e., the estimate (72.14b) holds true.

(3) Existence of p and well-posedness of (72.12). Since $\boldsymbol{f}$ and $\partial_t \boldsymbol{u}$ are in $L^2(J; \boldsymbol{L}^2(D))$, we can define the linear form $\boldsymbol{S} \in L^2(J; \boldsymbol{H}^{-1}(D)) = L^2(J; \boldsymbol{H}^1_0(D))'$ (see Lemma 64.20(i)) such that

$$\langle \boldsymbol{S}, \boldsymbol{w} \rangle_{L^2(\boldsymbol{H}^{-1}), L^2(\boldsymbol{H}^1_0)} = \int_J \big((\partial_t \boldsymbol{u}(t) - \boldsymbol{f}(t), \boldsymbol{w}(t))_{\boldsymbol{L}^2} + a(\boldsymbol{u}(t), \boldsymbol{w}(t)) \big) \mathrm{d}t,$$

for all $\boldsymbol{w} \in L^2(J; \boldsymbol{H}^1_0(D))$. Since $\boldsymbol{u}$ solves the constrained weak formulation (72.11), $\boldsymbol{S}$ annihilates $L^2(J; \boldsymbol{\mathcal{V}})$. Owing to Lemma 72.4, this implies that there exists $p \in L^2(J; L^2_*(D))$ such that the pair $(\boldsymbol{u}, p)$ solves the mixed weak formulation (72.12). Moreover, the inf-sup inequality (72.7) implies that

$$\beta \|p(t)\|_{L^2} \le \frac{\ell_D}{C_{\mathrm{KPS}}} \big(\|\partial_t \boldsymbol{u}(t)\|_{\boldsymbol{L}^2} + \|\boldsymbol{f}(t)\|_{\boldsymbol{L}^2} \big) + 2\mu \|\boldsymbol{u}(t)\|_{\boldsymbol{V}},$$

for a.e. $t \in J$. Squaring and using the definition of the time scale ρ yields

$$\beta^2 \|p(t)\|^2_{L^2} \le 4\rho\mu \big(\|\partial_t \boldsymbol{u}(t)\|^2_{\boldsymbol{L}^2} + \|\boldsymbol{f}(t)\|^2_{\boldsymbol{L}^2} \big) + 8\mu^2 \|\boldsymbol{u}(t)\|^2_{\boldsymbol{V}}.$$

Integrating over J, and using the above estimates on $\|\partial_t \boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}$ and on $\|\boldsymbol{u}\|_{L^2(J;\boldsymbol{V})}$ proves the estimate (72.14c) on the pressure. Finally, uniqueness of the solution to (72.12) follows from the a priori estimates. □

Lemma 72.4 (Space-time de Rham). *Let $\boldsymbol{S} \in L^2(J; \boldsymbol{H}^{-1}(D))$. Then, the linear form S satisfies $\langle \boldsymbol{S}, \boldsymbol{w} \rangle_{L^2(\boldsymbol{H}^{-1}), L^2(\boldsymbol{H}^1_0)} = 0$ for all $\boldsymbol{w} \in L^2(J; \boldsymbol{\mathcal{V}})$ iff there exists $p \in L^2(J; L^2_*(D))$ s.t. $\langle \boldsymbol{S}, \boldsymbol{w} \rangle_{L^2(\boldsymbol{H}^{-1}), L^2(\boldsymbol{H}^1_0)} = (p, \nabla\cdot \boldsymbol{w})_{L^2(J;L^2_*(D))}$ for all $\boldsymbol{w} \in L^2(J; \boldsymbol{H}^1_0(D))$.*

Proof. See Exercise 72.2. □

Remark 72.5 (Pressure gradient). Defining the gradient operator

$$\nabla : L^2(J; L^2_*(D)) \to L^2(J; \boldsymbol{H}^{-1}(D))$$

such that for all $q \in L^2(J; L^2_*(D))$,

$$\langle \nabla q, \boldsymbol{w} \rangle_{L^2(\boldsymbol{H}^{-1}), L^2(\boldsymbol{H}^1_0)} := -(q, \nabla\cdot\boldsymbol{w})_{L^2(J;L^2_*(D))},$$

for all $\boldsymbol{w} \in L^2(J; \boldsymbol{H}^1_0(D))$, Lemma 72.4 means that $\boldsymbol{S} \in L^2(J; \boldsymbol{H}^{-1}(D))$ annihilates $L^2(J; \boldsymbol{\mathcal{V}})$ iff there is $p \in L^2(J; L^2_*(D))$ s.t. $\boldsymbol{S} = -\nabla p$. □

Remark 72.6 ($\boldsymbol{\mathcal{V}}'$ vs. $\boldsymbol{V}'$). In §71.1 and, in particular, Lemma 71.2, we showed that the right test function to obtain an optimal estimate on the weak time derivative in $L^2(J; V')$ in the parabolic equation $\partial_t v + A(v) = f$ is $A^{-*}(\partial_t v)$. Let $\boldsymbol{A} : \boldsymbol{\mathcal{V}} \to \boldsymbol{\mathcal{V}}'$ be the *Stokes operator* defined by $\langle \boldsymbol{A}(\boldsymbol{v}), \boldsymbol{w} \rangle_{\boldsymbol{\mathcal{V}}',\boldsymbol{\mathcal{V}}} := a(\boldsymbol{v}, \boldsymbol{w})$ for all $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{\mathcal{V}}$. The operator $\boldsymbol{A}$ is bijective, self-adjoint, and its inverse is compact (recall that the embedding $\boldsymbol{\mathcal{V}} \subset \boldsymbol{H}^1(D) \hookrightarrow \boldsymbol{L}^2(D)$ is compact). The constrained weak formulation (72.11) consists of seeking $\boldsymbol{u} \in X(J; \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{V}}')$ so that $\partial_t \boldsymbol{u} + \boldsymbol{A}(\boldsymbol{u}) = \boldsymbol{f}$ in $L^2(J; \boldsymbol{\mathcal{V}}')$. Testing this equation by $\boldsymbol{A}^{-1}(\partial_t \boldsymbol{u})$ gives $\|\partial_t \boldsymbol{u}\|_{L^2(J;\boldsymbol{\mathcal{V}}')} \le c\|\boldsymbol{f}\|_{L^2(J;\boldsymbol{\mathcal{V}}')}$. So, a natural question that comes to mind is whether the norms $\|\cdot\|_{\boldsymbol{\mathcal{V}}'}$ and $\|\cdot\|_{\boldsymbol{V}'}$ are equivalent (recall that $\boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$). If it were the case, the above inequality would give us an estimate on $\|\partial_t \boldsymbol{u}\|_{L^2(J;\boldsymbol{V}')}$. It is clear that $\|\boldsymbol{v}\|_{\boldsymbol{\mathcal{V}}'} \le \|\boldsymbol{v}\|_{\boldsymbol{V}'}$ for all $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$, but unfortunately the converse is false; see Guermond [142, Thm. 4.1], Guermond and Salgado [161, Thm. 32] for counterexamples. In conclusion, when $\boldsymbol{f} \in L^2(J; \boldsymbol{V}')$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}}$, one only has $\boldsymbol{u} \in X(J; \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{V}}')$, and it is not possible to derive an a priori estimate on $\partial_t \boldsymbol{u}$ in $L^2(J; \boldsymbol{V}')$. □

72.4 Mixed weak formulation with rough data

In this section, we revisit the question of the regularity in time of the pressure and of the time derivative of the velocity by considering data with minimal regularity, i.e., $\boldsymbol{f} \in L^2(J; \boldsymbol{H}^{-1}(D))$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}} := \{\boldsymbol{v} \in \boldsymbol{L}^2(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D}\cdot\boldsymbol{n} = 0\}$. We will see that in this setting the notion of weak time derivative for the velocity in $L^2(J; \boldsymbol{H}^{-1}(D))$ is not sufficient and the pressure may not be in $L^2(J; L^2_*(D))$. As a result, we have to introduce distributional time derivatives to extend the notion of weak time derivatives.

We first introduce the notion of distributional time derivative. For every separable Hilbert space V, we define

$$H^1(J; V) := \{w \in L^2(J; V) \mid \partial_t w \in L^2(J; V)\}, \tag{72.18a}$$

$$H^1_0(J; V) := \{w \in H^1(J; V) \mid w(0) = 0,\ w(T) = 0\}. \tag{72.18b}$$

The definition of $H^1_0(J; V)$ is meaningful owing to Lemma 64.40. Notice that $H^1(J; V) = X(J; V, V)$. It can be shown that $H^1(J; V)$ is a Hilbert space when equipped with the inner product

$$(v, w)_{H^1(J;V)} := \int_J ((v(t), w(t))_V + T^2(\partial_t v(t), \partial_t w(t))_V)\mathrm{d}t$$

and that $H^1_0(J; V)$ is a closed subspace of $H^1(J; V)$. We denote the dual of $H^1_0(J; V)$ by $H^{-1}(J; V')$. For all $v \in L^2(J; V')$, we define the distributional time derivative $\hat{\partial}_t v$ to be the linear form in

$H^{-1}(J;V') := (H^1_0(J;V))'$ s.t. the following holds true for all $w \in H^1_0(J;V)$:

$$\langle \hat\partial_t v, w\rangle_{H^{-1}(V'),H^1_0(V)} := -\int_J \langle v, \partial_t w\rangle_{V',V}\mathrm{d}t, \tag{72.19}$$

where $H^{-1}(V')$ means $H^{-1}(J;V')$ and $H^1_0(V)$ means $H^1_0(J;V)$. The distributional time derivative is an extension of the weak time derivative, i.e., $\hat\partial_t v = \partial_t v$ for all $v \in H^1(J;V')$; see Exercise 72.4.

Theorem 72.7 (Pressure regularity). *Assume that $\boldsymbol{f} \in L^2(J;\boldsymbol{H}^{-1}(D))$ and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{H}}$. Let $\boldsymbol{u}$ solve* (72.11)*. Then there exists $p \in H^{-1}(J;L^2_*(D))$ such that for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{H}^1_0(D))$,*

$$\begin{aligned}\langle \hat\partial_t \boldsymbol{u}, \boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} + \int_J a(\boldsymbol{u}(t),\boldsymbol{w}(t))\mathrm{d}t - \langle p, \nabla\cdot\boldsymbol{w}\rangle_{H^{-1}(L^2),H^1_0(L^2)}\\ = \int_J \langle \boldsymbol{f}(t),\boldsymbol{w}(t)\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}^1_0}\mathrm{d}t.\end{aligned} \tag{72.20}$$

Proof. Since $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$ by assumption and $L^2(J;\boldsymbol{\mathcal{V}}) \hookrightarrow L^2(J;\boldsymbol{H}^{-1}(D))$, $\boldsymbol{u}$ has a distributional time derivative $\hat\partial_t \boldsymbol{u} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$ which satisfies

$$\langle \hat\partial_t \boldsymbol{u}, \boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} := -\int_J \langle \boldsymbol{u}, \partial_t \boldsymbol{w}\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}^1_0}\mathrm{d}t = -\int_J (\boldsymbol{u},\partial_t\boldsymbol{w})_{\boldsymbol{L}^2}\mathrm{d}t,$$

for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{H}^1_0(D))$. Moreover, the weak time derivative $\partial_t \boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}}')$ satisfies for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{\mathcal{V}}) \subset L^2(J;\boldsymbol{\mathcal{V}})$,

$$\langle \partial_t \boldsymbol{u}, \boldsymbol{w}\rangle_{L^2(\boldsymbol{\mathcal{V}}');L^2(\boldsymbol{\mathcal{V}})} = \int_J \langle \partial_t \boldsymbol{u}, \boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}}',\boldsymbol{\mathcal{V}}}\mathrm{d}t = -\int_J (\boldsymbol{u},\partial_t\boldsymbol{w})_{\boldsymbol{L}^2}\mathrm{d}t,$$

owing to the integration by parts formula from Lemma 64.40, since $\boldsymbol{u},\boldsymbol{w} \in X(J;\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{V}}')$, $\boldsymbol{w}(0) = \boldsymbol{w}(T) = 0$, and $\langle \boldsymbol{u}, \partial_t\boldsymbol{w}\rangle_{\boldsymbol{\mathcal{V}},\boldsymbol{\mathcal{V}}'} = (\boldsymbol{u},\partial_t\boldsymbol{w})_{\boldsymbol{L}^2}$. Thus, $\hat\partial_t\boldsymbol{u} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$ and $\partial_t\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}}')$ coincide on $H^1_0(J;\boldsymbol{\mathcal{V}})$. Consider now the linear form $\boldsymbol{S} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$ such that for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{H}^1_0(D))$,

$$\begin{aligned}&\langle \boldsymbol{S},\boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} :=\\ &\qquad \langle \hat\partial_t\boldsymbol{u},\boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} + \int_J \big(a(\boldsymbol{u}(t),\boldsymbol{w}(t)) - \langle \boldsymbol{f}(t),\boldsymbol{w}(t)\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}^1_0}\big)\mathrm{d}t.\end{aligned}$$

Since $\boldsymbol{u}$ solves (72.11), the identity on the time derivatives shows that $\langle \boldsymbol{S},\boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} = 0$ for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{\mathcal{V}})$. Owing to Lemma 72.8, there is $p \in H^{-1}(J;L^2_*(D))$ so that (72.20) holds true. □

Lemma 72.8 (Space-time de Rham in H^{-1}). *Let $\boldsymbol{S} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$. Then, the linear form S satisfies $\langle \boldsymbol{S},\boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} = 0$ for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{\mathcal{V}})$ iff there is $p \in H^{-1}(J;L^2_*(D))$ s.t. $\langle \boldsymbol{S},\boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} = \langle p, \nabla\cdot\boldsymbol{w}\rangle_{H^{-1}(L^2),H^1_0(L^2)}$ for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{H}^1_0(D))$.*

Proof. See Exercise 72.5. □

Remark 72.9 (Pressure gradient). For all $q \in H^{-1}(J;L^2_*(D))$ (which is by definition the dual space of $H^1_0(J,L^2_*(D))$), we define

$$\hat\nabla q \in H^{-1}(J;\boldsymbol{H}^{-1}(D)) = (H^1_0(J;\boldsymbol{H}^1_0(D)))',$$

so that the following holds true for all $\boldsymbol{w} \in H^1_0(J;\boldsymbol{H}^1_0(D))$:

$$\langle \hat{\nabla} q, \boldsymbol{w}\rangle_{H^{-1}(\boldsymbol{H}^{-1}),H^1_0(\boldsymbol{H}^1_0)} = -\langle q, \nabla\cdot\boldsymbol{w}\rangle_{H^{-1}(L^2_*),H^1_0(L^2_*)}.$$

Lemma 72.8 means that $\boldsymbol{S} \in H^{-1}(J;\boldsymbol{H}^{-1}(D))$ annihilates $H^1_0(J;\boldsymbol{\mathcal{V}})$ iff there is $p \in H^{-1}(J;L^2_*(D))$ s.t. $\boldsymbol{S} = -\hat{\nabla} p$. Moreover, (72.20) can be rewritten $\hat{\partial}_t \boldsymbol{u} - \nabla\cdot\mathbb{s}(\boldsymbol{u}) + \hat{\nabla} p = \boldsymbol{f}$ in $H^{-1}(J;\boldsymbol{H}^{-1}(D))$. Notice that only $\nabla\cdot\mathbb{s}(\boldsymbol{u})$ and $\boldsymbol{f}$ are in $L^2(J;\boldsymbol{H}^{-1}(D))$. The sum $\hat{\partial}_t\boldsymbol{u} + \hat{\nabla} p$ is in $L^2(J;\boldsymbol{H}^{-1}(D))$, but this may not be the case of the terms $\hat{\partial}_t\boldsymbol{u}$ and $\hat{\nabla} p$ taken individually. □

Remark 72.10 (Finer regularity results). Using a Fourier technique and assuming only $\boldsymbol{f} \in L^2(J;\boldsymbol{H}^{-1}(D))$, it is possible to prove $\boldsymbol{u} \in H^{\frac{1}{2}-\epsilon}(J;\boldsymbol{L}^2(D))$ and $p \in H^{-\frac{1}{2}-\epsilon}(J;L^2_*(D))$ for all $\epsilon > 0$; see Lions [217], [219, I§6.5]. Furthermore, let $q \in (1,\infty)$, $r \in (1,\infty)$, and $\epsilon > 0$. Assume that the Laplace operator and the Stokes operator, both with homogeneous Dirichlet boundary condition, are isomorphisms between $\boldsymbol{W}^{2,q}(D) \cap \boldsymbol{W}^{1,q}_0(D)$ and $\boldsymbol{L}^q(D)$ and between $\boldsymbol{\mathcal{W}}^{2,q} := \boldsymbol{W}^{2,q}(D) \cap \boldsymbol{W}^{1,q}_0(D) \cap \boldsymbol{\mathcal{H}}^q$ and $\boldsymbol{\mathcal{H}}^q := \{\boldsymbol{v} \in \boldsymbol{L}^q(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D}\cdot\boldsymbol{n} = 0\}$, respectively (these properties hold true if D is either convex or ∂D is of class C^1). Then for all $\boldsymbol{f} \in L^r(J;\boldsymbol{L}^q(D))$ and all $\boldsymbol{u}_0 \in \boldsymbol{W}^{1-\frac{1}{r}+\epsilon,q}(D) \cap \boldsymbol{\mathcal{H}}^q$, the time-dependent Stokes problem (72.1) has a unique solution with $\boldsymbol{u} \in L^r(J;\boldsymbol{\mathcal{W}}^{2,q})$, $\partial_t\boldsymbol{u} \in L^r(J;\boldsymbol{\mathcal{H}}^q)$, $p \in L^r(J;L^q_*(D))$, $\nabla p \in L^r(J;\boldsymbol{L}^q(D))$; see Sohr and von Wahl [264, Thm. 2.12]. □

Exercises

Exercise 72.1 (Non-homogeneous Dirichlet condition). Consider the time-dependent Stokes equations (72.1) with the non-homogeneous Dirichlet condition $\boldsymbol{u} = \boldsymbol{g}$ enforced over the whole boundary ∂D for all $t \in J$. Assume that $\int_{\partial D} \boldsymbol{g}\cdot\boldsymbol{n} = 0$ for all $t \in J$. Assume that the data $\boldsymbol{f}$ and $\boldsymbol{g}$ are smooth so that the solution $(\boldsymbol{u}, p)$ is smooth. Assume that there is a smooth lifting $\boldsymbol{u}_{\boldsymbol{g}}$ of the boundary datum so that $\boldsymbol{u}_{\boldsymbol{g}}\cdot\boldsymbol{n} = \boldsymbol{g}$ on $\partial D \times J$ and $\nabla\cdot\boldsymbol{u}_{\boldsymbol{g}} = 0$ on $D \times J$. (i) Write the equations satisfied by $\boldsymbol{u}_0 := \boldsymbol{u} - \boldsymbol{u}_{\boldsymbol{g}}$. (ii) Verify that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}^2} = (\boldsymbol{f},\boldsymbol{u}_0)_{\boldsymbol{L}^2} - (\partial_t\boldsymbol{u}_{\boldsymbol{g}},\boldsymbol{u}_0)_{\boldsymbol{L}^2} - 2\mu(\mathbb{e}(\boldsymbol{u}_{\boldsymbol{g}}),\mathbb{e}(\boldsymbol{u}_0))_{\mathbb{L}^2}.$$

(iii) Establish a priori bound on $\boldsymbol{u}_0$ of the form $\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2} + 2\mu\|\mathbb{e}(\boldsymbol{u}_0)\|^2_{\mathbb{L}^2} \le \Phi(T,\boldsymbol{f},\boldsymbol{u}_{\boldsymbol{g}}) + \frac{1}{T}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}$.

Exercise 72.2 (Space-time de Rham in L^2). (i) Show that the operator $\nabla\cdot : L^2(J;\boldsymbol{H}^1_0(D)) \to L^2(J;L^2_*(D))$ is surjective. (*Hint*: invoke Lemma 53.9, Lemma C.44, and Corollary 64.14.) (ii) Show that $\boldsymbol{S} \in L^2(J;\boldsymbol{H}^{-1}(D))$ satisfies $\int_J \langle \boldsymbol{S},\boldsymbol{w}\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}^1_0}\mathrm{d}t = 0$ for all $\boldsymbol{w} \in L^2(J;\boldsymbol{\mathcal{V}})$ iff there is $p \in L^2(J;L^2_*(D))$ s.t. $\int_J \langle \boldsymbol{S},\boldsymbol{w}\rangle_{\boldsymbol{H}^{-1},\boldsymbol{H}^1_0}\mathrm{d}t = \int_J (p,\nabla\cdot\boldsymbol{w})_{L^2}\mathrm{d}t$ for all $\boldsymbol{w} \in L^2(J;\boldsymbol{H}^1_0(D))$. (*Hint*: use the closed range theorem.)

Exercise 72.3 (Variable viscosity). Assume that μ depends on $\boldsymbol{x} \in D$, and set $0 < \mu_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D}\mu$, $\mu_\sharp := \operatorname{ess\,sup}_{\boldsymbol{x}\in D}\mu < \infty$. Consider the mixed weak formulation (72.12). Prove that $\mu_\flat\|\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{\mathcal{V}})} \le \frac{1}{4}\rho\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + \frac{1}{2}\|\boldsymbol{u}_0\|^2_{\boldsymbol{L}^2}$ with $\rho := C^{-2}_{\mathrm{KPS}}\frac{\ell_D^2}{\mu_\flat}$, $\|\partial_t\boldsymbol{u}\|^2_{L^2(J;\boldsymbol{L}^2)} \le \|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + 2\mu_\sharp\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}$, and $\|p\|^2_{L^2(J;L^2)} \le \frac{1}{\beta^2}\big(c_1\|\boldsymbol{f}\|^2_{L^2(J;\boldsymbol{L}^2)} + c_2\|\boldsymbol{u}_0\|^2_{\boldsymbol{\mathcal{V}}}\big)$ with $c_1 := \rho\mu_\flat(8+2\xi_\mu^2)$, $c_2 := \rho\mu_\flat\mu_\sharp(8+4\xi_\mu)$, and $\xi_\mu := \frac{\mu_\sharp}{\mu_\flat}$. (*Hint*: adapt the proof of Theorem 72.3.)

Exercise 72.4 (Distributional time derivative). Let $V \hookrightarrow L \equiv L' \hookrightarrow V'$ be a Gelfand triple. (i) Let $v \in X(J;V,V')$. Show that the action of $\hat{\partial}_t v \in H^{-1}(J;V')$ and of $\partial_t v \in L^2(J;V')$

coincide on $H^1_0(J;V)$. (*Hint*: use the integration by parts formula from Lemma 64.40.) (ii) Let $v \in H^1(J;L)$. Show that the action of $\hat{\partial}_t v \in H^{-1}(J;V')$ and of $\partial_t v \in L^2(J;L)$ coincide on $H^1_0(J;V)$. (*Hint*: as above.)

Exercise 72.5 (Space-time de Rham in H^{-1}). (i) Show that the operator $\nabla\cdot : H^1(J;\boldsymbol{H}^1_0(D)) \to H^1(J;L^2_*(D))$ is surjective. (*Hint*: proceed as in Exercise 72.2 and use Lemma 64.34.) (ii) Show that $\nabla\cdot : H^1_0(J;\boldsymbol{H}^1_0(D)) \to H^1_0(J;L^2_*(D))$ is surjective. (*Hint*: use Step (i) and Lemma 64.37.) (iii) Prove Lemma 72.8. (*Hint*: use the closed range theorem.)

Chapter 73

Monolithic time discretization

The present chapter deals with the approximation of the time-dependent Stokes equations. We use stable mixed finite elements for the space discretization in a conforming setting. The time discretization can be done with any of the techniques considered for the heat equation. For brevity, we focus on the implicit Euler scheme and on higher-order implicit Runge–Kutta (IRK) schemes. The discretization process gives at each time step a saddle point problem coupling the velocity and the pressure, so that the linear algebra is in general more involved than when dealing with the heat equation. Fractional-step methods based on a sequential computation of the velocity and the pressure are discussed in the next two chapters.

73.1 Model problem

In this chapter and the following two chapters, we consider the mixed weak formulation (72.12), i.e., we assume that homogeneous Dirichlet conditions are enforced on the velocity over the whole boundary, $\boldsymbol{f} \in L^2(J;\boldsymbol{L}^2(D))$, and $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{V}}$, where

$$\boldsymbol{\mathcal{V}} := \{\boldsymbol{v} \in \boldsymbol{V} \mid \nabla\cdot\boldsymbol{v} = 0\}, \qquad \boldsymbol{V} := \boldsymbol{H}^1_0(D). \tag{73.1}$$

The solution to (72.12) satisfies $\boldsymbol{u} \in X(J;\boldsymbol{\mathcal{V}},\boldsymbol{L}^2(D))$, i.e., $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$ and $\partial_t\boldsymbol{u} \in L^2(J;\boldsymbol{L}^2(D))$, and $p \in L^2(J;Q)$ with $Q := L^2_*(D)$. The weak formulation is as follows:

$$\begin{cases} (\partial_t\boldsymbol{u}(t),\boldsymbol{w})_{\boldsymbol{L}^2} + a(\boldsymbol{u}(t),\boldsymbol{w}) + b(\boldsymbol{w},p(t)) = (\boldsymbol{f}(t),\boldsymbol{w})_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}(t),q) = 0, \end{cases} \tag{73.2}$$

for all $\boldsymbol{w} \in \boldsymbol{V} := \boldsymbol{H}^1_0(D)$ and all $q \in Q := L^2_*(D)$, where the two equalities are understood to hold in $L^2(J)$. Notice that the second equation in (73.2) means that $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$. Recall that $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2}$ for all $\boldsymbol{v} \in \boldsymbol{V}$, and that Q is equipped with the L^2-norm.

We henceforth assume that there is some regularity pickup for the steady Stokes problem (53.1), i.e., there are real numbers c_{smo} and $s \in (0,1]$ such that the solution to the steady-state Stokes problem with source $\boldsymbol{s} \in \boldsymbol{L}^2(D)$, say $(\boldsymbol{\zeta}(\boldsymbol{s}),\theta(\boldsymbol{s})) \in \boldsymbol{V}\times Q$, is such that

$$\mu\ell_D^{-1}\|\boldsymbol{\zeta}(\boldsymbol{s})\|_{\boldsymbol{H}^{1+s}(D)} + \|\theta(\boldsymbol{s})\|_{H^s(D)} \le c_{\mathrm{smo}}\ell_D\|\boldsymbol{s}\|_{\boldsymbol{L}^2(D)}, \tag{73.3}$$

where ℓ_D is some characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$.

73.2 Space semi-discretization

In this section, we discuss the space discretization of (73.2) using conforming mixed finite elements and we perform the error analysis.

73.2.1 Discrete formulation

We adopt the same discrete setting as in §53.3. We assume that D is a polyhedron in $\mathbb{R}^d$ and $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is a shape-regular sequence of matching meshes so that each mesh covers D exactly. Let $(\boldsymbol{V}_h)_{h\in\mathcal{H}}$ and $(Q_h)_{h\in\mathcal{H}}$ be sequences of finite-dimensional spaces built using $(\mathcal{T}_h)_{h\in\mathcal{H}}$. We assume that the approximation setting is conforming, i.e., $\boldsymbol{V}_h \subset \boldsymbol{V} := \boldsymbol{H}^1_0(D)$ and $Q_h \subset Q := L^2_*(D)$ for all $h \in \mathcal{H}$ (this means in particular that the velocity boundary conditions are strongly enforced). We assume that the pairs $(\boldsymbol{V}_h, Q_h)_{h\in\mathcal{H}}$ are uniformly compatible, i.e., there exists a constant $\beta > 0$ such that for all $h \in \mathcal{H}$,

$$\inf_{q_h\in Q_h} \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_h} \frac{\int_D q_h \nabla\cdot\boldsymbol{v}_h \,\mathrm{d}x}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}\|q_h\|_{L^2(D)}} \geq \beta. \tag{73.4}$$

Let us set

$$\boldsymbol{\mathcal{V}}_h := \{\boldsymbol{v}_h \in \boldsymbol{V}_h \mid b(\boldsymbol{v}_h, q_h) = 0,\ \forall q_h \in Q_h\}. \tag{73.5}$$

Recall that $\boldsymbol{\mathcal{V}}_h$ is not in general a subspace of $\boldsymbol{\mathcal{V}}$. The discretization is said to be *well-balanced* when $\boldsymbol{\mathcal{V}}_h \subset \boldsymbol{\mathcal{V}}$ (see Remark 53.22).

Let $\boldsymbol{u}_h(0) \in \boldsymbol{\mathcal{V}}_h$ be uniquely defined by requiring that $a(\boldsymbol{u}_h(0), \boldsymbol{w}_h) = a(\boldsymbol{u}_0, \boldsymbol{w}_h)$ for all $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$ (recall that a is coercive on $\boldsymbol{V}_h$ and thus on $\boldsymbol{\mathcal{V}}_h$). The space semi-discrete problem is as follows: Find $\boldsymbol{u}_h \in H^1(J;\boldsymbol{V}_h)$ and $p_h \in L^2(J;Q_h)$ such that the following holds true in $L^2(J)$ for all $\boldsymbol{w}_h \in \boldsymbol{V}_h$ and all $q_h \in Q_h$:

$$\begin{cases} (\partial_t \boldsymbol{u}_h(t), \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{u}_h(t), \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h(t)) = (\boldsymbol{f}(t), \boldsymbol{w}_h)_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}_h(t), q_h) = 0. \end{cases} \tag{73.6}$$

Notice that the second equation in (73.6) implies that $\boldsymbol{u}_h \in H^1(J;\boldsymbol{\mathcal{V}}_h)$.

Proposition 73.1 (Well-posedness). *The discrete problem* (73.6) *is well-posed.*

Proof. See Exercise 73.1. □

73.2.2 Error equations and approximation operators

To gain some insight into the derivation of the error estimates, let us consider some discrete functions $\boldsymbol{v}_h \in H^1(J;\boldsymbol{\mathcal{V}}_h)$ and $q_h \in L^2(J;Q_h)$, and let us consider the following error decompositions for all $t \in J$:

$$\boldsymbol{e}_h(t) := \boldsymbol{u}_h(t) - \boldsymbol{v}_h(t), \qquad \boldsymbol{\eta}(t) := \boldsymbol{u}(t) - \boldsymbol{v}_h(t), \tag{73.7a}$$

$$\delta_h(t) := p_h(t) - q_h(t), \qquad \zeta(t) := p(t) - q_h(t). \tag{73.7b}$$

Notice that $\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$. Moreover, subtracting (73.6) from (73.2) and using the conformity of the approximation setting, we infer that for all $t \in J$ and all $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$,

$$(\partial_t \boldsymbol{e}_h, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{e}_h, \boldsymbol{w}_h) = (\partial_t \boldsymbol{\eta}, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{\eta}, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, \zeta). \tag{73.8}$$

Notice that $b(\boldsymbol{w}_h, \delta_h) = 0$ whenever $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$.

Taking inspiration from the error analysis for the heat equation (see §66.3), a natural way to proceed is to consider the counterpart of the elliptic projection introduced for parabolic problems. For the time-dependent Stokes equations, we define the operators $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}} : \boldsymbol{V}\times Q \to \boldsymbol{V}_h$ and $\mathcal{S}_h^{\mathrm{p}} : \boldsymbol{V}\times Q \to Q_h$ such that for all $(\boldsymbol{v},q) \in \boldsymbol{V}\times Q$, $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q) \in \boldsymbol{V}_h$ and $\mathcal{S}_h^{\mathrm{p}}(\boldsymbol{v},p) \in Q_h$ are defined as follows: For all $(\boldsymbol{w}_h,r_h) \in \boldsymbol{V}_h\times Q_h$:

$$a(\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q),\boldsymbol{w}_h) + b(\boldsymbol{w}_h,\mathcal{S}_h^{\mathrm{p}}(\boldsymbol{v},q)) := a(\boldsymbol{v},\boldsymbol{w}_h) + b(\boldsymbol{w}_h,q), \tag{73.9a}$$

$$b(\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q),r_h) := b(\boldsymbol{v},r_h). \tag{73.9b}$$

Notice that $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q) \in \boldsymbol{\mathcal{V}}_h$ whenever $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$. Then setting $\boldsymbol{e}_h(t) := \boldsymbol{u}_h(t) - \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}(t),p(t))$ and $\boldsymbol{\eta}(t) := \boldsymbol{u}(t) - \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}(t),p(t))$ for all $t \in J$ and all $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$, we observe that $\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$ and that the error equation (73.8) becomes

$$(\partial_t \boldsymbol{e}_h,\boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{e}_h,\boldsymbol{w}_h) = (\partial_t\boldsymbol{\eta},\boldsymbol{w}_h)_{\boldsymbol{L}^2}. \tag{73.10}$$

The error equation (73.10) shows that we need to measure the approximation properties of $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}$ to derive a velocity error estimate. But this is precisely what has been done in Chapter 53 in the context of the steady Stokes equations. Indeed, Theorem 53.17 and Theorem 53.19 (see also Corollary 50.5 and Remark 50.6 for a more abstract setting) show that there is c s.t. for all $(\boldsymbol{v},q) \in \boldsymbol{V}\times Q$ and all $h \in \mathcal{H}$,

$$\begin{aligned}
\|\boldsymbol{v} - \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q)\|_{\boldsymbol{V}} &\le c\Big(\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{\boldsymbol{V}} + \frac{1}{\mu}\inf_{q_h\in Q_h}\|q-q_h\|_{L^2}\Big),\\
\|q - \mathcal{S}_h^{\mathrm{p}}(\boldsymbol{v},q)\|_{L^2} &\le c\Big(\mu\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{\boldsymbol{V}} + \inf_{q_h\in Q_h}\|q-q_h\|_{L^2}\Big),\\
\|\boldsymbol{v} - \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},q)\|_{\boldsymbol{L}^2} &\le c\,h^s\ell_D^{1-s}\Big(\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{\boldsymbol{V}} + \frac{1}{\mu}\inf_{q_h\in Q_h}\|q-q_h\|_{L^2}\Big).
\end{aligned}$$

One possible drawback of using the approximation operator $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}$ to estimate the velocity error is that the field $\boldsymbol{\eta}$ depends on the pressure. An alternative is to consider the projection operator $\boldsymbol{P}_h^{\mathrm{s}} : \boldsymbol{V} \to \boldsymbol{\mathcal{V}}_h$ such that

$$a(\boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{v}),\boldsymbol{w}_h) = a(\boldsymbol{v},\boldsymbol{w}_h), \qquad \forall(\boldsymbol{v},\boldsymbol{w}_h) \in \boldsymbol{V}\times\boldsymbol{\mathcal{V}}_h. \tag{73.11}$$

Notice that $\boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{v}) = \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{v},0)$ for all $\boldsymbol{v} \in \boldsymbol{V}$. Then setting $\boldsymbol{e}_h(t) := \boldsymbol{u}_h(t) - \boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{u}(t))$ and $\boldsymbol{\eta}(t) := \boldsymbol{u}(t) - \boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{u}(t))$ for all $t \in J$, we observe that $\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$ and that (73.8) now becomes

$$(\partial_t \boldsymbol{e}_h,\boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{e}_h,\boldsymbol{w}_h) = (\partial_t\boldsymbol{\eta},\boldsymbol{w}_h)_{\boldsymbol{L}^2} + b(\boldsymbol{w}_h,p-q_h), \tag{73.12}$$

for all $q_h \in H^1(J;Q_h)$. This shows that the velocity error estimate is still dependent on the pressure approximation, but at least the dependence on the viscosity can be avoided. This topic is further discussed below. For the time being, we recall from Lemma 53.20 that the projection operator $\boldsymbol{P}_h^{\mathrm{s}}$ enjoys optimal approximation properties. Indeed, provided the inf-sup condition (73.4) is satisfied, the following holds true for all $\boldsymbol{v} \in \boldsymbol{V}$ and any Fortin operator $\boldsymbol{\Pi}_h \in \mathcal{L}(\boldsymbol{V};\boldsymbol{V}_h)$:

$$\|\boldsymbol{v} - \boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{v})\|_{\boldsymbol{V}} \le \tilde{c}_{1h}\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h}\|\boldsymbol{v}-\boldsymbol{v}_h\|_{\boldsymbol{V}}, \tag{73.13}$$

with $\tilde{c}_{1h} := \frac{\|a\|}{\alpha}(1+\|\boldsymbol{\Pi}_h\|_{\mathcal{L}(\boldsymbol{V};\boldsymbol{V}_h)})$. Notice that the ratio $\frac{\|a\|}{\alpha}$ is independent of the viscosity and only depends on the constant C_{K} from Korn's inequality, whereas $\|\boldsymbol{\Pi}_h\|_{\mathcal{L}(\boldsymbol{V};\boldsymbol{V}_h)}$ can be bounded by $\frac{\|b\|}{\beta}$ (see Lemma 26.9).

Remark 73.2 (Initialization). The initialization for the space semi-discrete problem (73.6) can be rewritten $\boldsymbol{u}_h(0) := \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{u}_0) = \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}_0, 0)$. □

Remark 73.3 ($\boldsymbol{L}^2$-orthogonal projection). Yet another error equation can be derived if one considers the error decomposition using the $\boldsymbol{L}^2$-orthogonal projection from $\boldsymbol{V}$ onto $\boldsymbol{\mathcal{V}}_h$. This choice eliminates the term $(\partial_t\boldsymbol{\eta}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}$ from the right-hand side of (73.8). However, strong assumptions on the mesh sequence are required to obtain optimal approximation properties in the $\|\cdot\|_{\boldsymbol{V}}$-norm for the $\boldsymbol{L}^2$-orthogonal projection (see Proposition 22.21 and Remark 22.23.) □

73.2.3 Error analysis

We are now ready to perform the error analysis of the semi-discrete problem (73.6). We start with the natural approach where we use the approximation operators $(\boldsymbol{\mathcal{S}}_h^{\mathrm{V}}, \mathcal{S}_h^{\mathrm{P}})$ defined in (73.9).

Theorem 73.4 (Error estimates). *Let $(\boldsymbol{u}, p)$ solve (73.2) and assume that $\boldsymbol{u} \in H^1(J;\boldsymbol{V})$ and $p \in H^1(J;L^2_*(D))$. Let $(\boldsymbol{u}_h, p_h)$ solve (73.6). Let $\boldsymbol{\eta} := \boldsymbol{u} - \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}, p)$, $\zeta := p - \mathcal{S}_h^{\mathrm{P}}(\boldsymbol{u}, p)$ for all $t \in J$, and let $\boldsymbol{e}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{0}, p(0))$.* (i) *The following holds true for all $h \in \mathcal{H}$:*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{L^2(J;\boldsymbol{V})} \le \|\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{V})} + c_1\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)} + c_2\|\boldsymbol{e}_h^0\|_{\boldsymbol{L}^2}, \tag{73.14a}$$

$$\|p - p_h\|_{L^2(J;L^2)} \le \|\zeta\|_{L^2(J;L^2)} + c_3\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)} + c_4\|\boldsymbol{e}_h^0\|_{\boldsymbol{V}}, \tag{73.14b}$$

with $c_1 := \frac{\sqrt{\rho}}{2\sqrt{\mu}}$, $c_2 := \frac{1}{\sqrt{2\mu}}$, $c_3 := \frac{1}{\beta}\sqrt{12\rho\mu}$, $c_4 := \frac{1}{\beta}\sqrt{10\rho}\mu$, and the time scale $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$. (ii) *We have for all $t \in (0,T]$ with $J_t := (0,t)$,*

$$\|(\boldsymbol{u} - \boldsymbol{u}_h)(t)\|_{\boldsymbol{L}^2} \le \|\boldsymbol{\eta}(t)\|_{\boldsymbol{L}^2} + \frac{\sqrt{\rho}}{\sqrt{2}}\|e^{-\frac{t-\cdot}{\rho}}\partial_t\boldsymbol{\eta}\|_{L^2(J_t;\boldsymbol{L}^2)} + e^{-\frac{t}{\rho}}\|\boldsymbol{\eta}_0\|_{\boldsymbol{L}^2}.$$

Proof. (1) Let us set $\boldsymbol{e}_h := \boldsymbol{u}_h - \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}, p)$, $\boldsymbol{\eta} := \boldsymbol{u} - \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}, p)$, and $\delta_h := p_h - \mathcal{S}_h^{\mathrm{P}}(\boldsymbol{u}, p)$ for all $t \in J$ (these quantities are well defined since $\boldsymbol{u} \in H^1(J;\boldsymbol{V})$ and $p \in H^1(J;L^2_*(D))$). Notice that $\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$, and proceeding as in the derivation of (73.10), we have for all $t \in J$ and all $\boldsymbol{w}_h \in \boldsymbol{V}_h$,

$$(\partial_t\boldsymbol{e}_h, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{e}_h, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, \delta_h) = (\partial_t\boldsymbol{\eta}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}. \tag{73.15}$$

Using the test function $\boldsymbol{w}_h := \boldsymbol{e}_h(t)$ in (73.15) for all $t \in J$ and using that $b(\boldsymbol{e}_h(t), \delta_h) = 0$ since $\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$, together with the coercivity of a and Young's inequality gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2 + 2\mu\|\boldsymbol{e}_h\|_{\boldsymbol{V}}^2 \le \frac{1}{4\mu}\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{V}'}^2 + \mu\|\boldsymbol{e}_h\|_{\boldsymbol{V}}^2.$$

Since $\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{V}'} \le C_{\mathrm{KPS}}^{-1}\ell_D\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{L}^2}$ owing to (72.4), recalling the definition of ρ, integrating over $t \in J$, and dropping the nonnegative term $\|\boldsymbol{e}_h(T)\|_{\boldsymbol{L}^2}^2$ from the left-hand side yields

$$2\mu\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}^2 \le \frac{1}{2}\rho\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2 + \|\boldsymbol{e}_h(0)\|_{\boldsymbol{L}^2}^2.$$

Invoking the triangle inequality yields (73.14a) since $\boldsymbol{e}_h(0) = \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}_0, 0) - \boldsymbol{\mathcal{S}}_h^{\mathrm{V}}(\boldsymbol{u}_0, p(0)) =: -\boldsymbol{e}_h^0$ owing to the linearity of $\boldsymbol{\mathcal{S}}_h^{\mathrm{V}}$. Note that the above arguments are the same as those used for the heat equation (see Lemma 65.10) and for the a priori estimate on the solution to the time-dependent Stokes problem (see Theorem 72.3).

(2) To derive the pressure error, we first bound $\|\partial_t\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{L}^2)}$ and then invoke the inf-sup condition (73.4). Proceeding as in the proof of (72.17), we now use the test function $\boldsymbol{w}_h = \partial_t\boldsymbol{e}_h(t)$ for all $t \in J$ in (73.15). Since $\partial_t\boldsymbol{e}_h(t) \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$, we infer that

$$\|\partial_t\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2 + \frac{\mathrm{d}}{\mathrm{d}t}a(\boldsymbol{e}_h, \boldsymbol{e}_h) \le \frac{1}{2}\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\|\partial_t\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2.$$

Integrating over $t \in J$, using the coercivity and the boundedness of a, and dropping the nonnegative term $2\mu\|\boldsymbol{e}_h(T)\|_{\boldsymbol{V}}^2$ from the left-hand side, we infer that

$$\|\partial_t \boldsymbol{e}_h\|_{L^2(J;\boldsymbol{L}^2)}^2 \le \|\partial_t \boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2 + 2\mu\|\boldsymbol{e}_h(0)\|_{\boldsymbol{V}}^2.$$

We can now invoke the inf-sup condition (73.4) and use (73.15) for all $\boldsymbol{w}_h \in \boldsymbol{V}_h$ to infer that for all $t \in J$,

$$\begin{aligned}\beta\|\delta_h\|_{L^2} &\le \sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h} \frac{|b(\boldsymbol{w}_h,\delta_h)|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}}}\\ &= \sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h} \frac{|(\partial_t\boldsymbol{e}_h,\boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{e}_h,\boldsymbol{w}_h) - (\partial_t\boldsymbol{\eta},\boldsymbol{w}_h)_{\boldsymbol{L}^2}|}{\|\boldsymbol{w}_h\|_{\boldsymbol{V}}}\\ &\le \sqrt{\rho\mu}\big(\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{L}^2} + \|\partial_t\boldsymbol{e}_h\|_{\boldsymbol{L}^2}\big) + 2\mu\|\boldsymbol{e}_h\|_{\boldsymbol{V}},\end{aligned}$$

where we used the triangle inequality, (72.4), the definition of ρ, and the boundedness of a. Squaring, integrating over $t \in J$, and using the above bound on $\|\partial_t\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{L}^2)}^2$ gives

$$\begin{aligned}\beta^2\|\delta_h\|_{L^2(J;L^2)}^2 &\le 4\rho\mu\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2 + 4\rho\mu\|\partial_t\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{L}^2)}^2 + 8\mu^2\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}^2\\ &\le 8\rho\mu\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2 + 8\rho\mu^2\|\boldsymbol{e}_h(0)\|_{\boldsymbol{V}}^2 + 8\mu^2\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}^2.\end{aligned}$$

Invoking the bound on $\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}^2$ from Step (1) and using that $\|\boldsymbol{e}_h(0)\|_{\boldsymbol{L}^2}^2 \le \rho\mu\|\boldsymbol{e}_h(0)\|_{\boldsymbol{V}}^2$, we infer that

$$\beta^2\|\delta_h\|_{L^2(J;L^2)}^2 \le 10\rho\mu\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2 + 12\rho\mu^2\|\boldsymbol{e}_h(0)\|_{\boldsymbol{V}}^2.$$

This yields (73.14b) after taking the square root and invoking the triangle inequality. Notice that the above arguments are the same as those invoked to derive the a priori estimate (72.14c) on the pressure.

(3) The third error estimate is obtained by using the test function $\boldsymbol{w}_h := e^{2\frac{t}{\rho}}\boldsymbol{e}_h(t)$ in (73.15) for all $t \in J$, proceeding as in the proof of Theorem 66.9 since we still have $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$ for all $t \in J$, and additionally invoking as above that $\|\partial_t\boldsymbol{\eta}(t)\|_{\boldsymbol{V}'} \le \sqrt{\rho\mu}\|\partial_t\boldsymbol{\eta}(t)\|_{\boldsymbol{L}^2}$ for all $t \in J$. □

Remark 73.5 (Convergence rates). Let $r \in [1,k]$, where $k \ge 1$ is the polynomial degree of the finite elements used to build $\boldsymbol{V}_h$, and let $k' \ge k-1$ be the polynomial degree of the finite elements used to build Q_h. Assume that $\boldsymbol{u} \in X(J;\boldsymbol{H}^{r+1}(D),\boldsymbol{H}^r(D))$, $p \in X(J;H^{r+1}(D),H^r(D))$, $\boldsymbol{u}_0 \in \boldsymbol{H}^{r+1}(D)$, and $p(0) \in H^r(D)$. The bounds from Theorem 73.4 imply that the error on the velocity in the $L^2(J;\boldsymbol{H}^1(D))$-norm and the error on the pressure in the $L^2(J;L^2(D))$-norm decay as $\mathcal{O}(h^r)$. Moreover, the error on the velocity in the $C^0(\overline{J};\boldsymbol{L}^2(D))$-norm decays as $\mathcal{O}(h^{r+s})$, where $s \in (0,1]$ is the regularity pickup index ($s = 1$ if there is full regularity pickup). Just like parabolic equations, the error induced by approximating the initial data converges to zero exponentially as T grows. □

The velocity error estimate derived in Theorem 73.4 may not be sharp whenever the space discretization scheme is not well-balanced, i.e., whenever $\boldsymbol{\mathcal{V}}_h \not\subset \boldsymbol{\mathcal{V}}$. In this situation, the error induced by the approximation operator $\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}$ can be dominated by the approximation error on the pressure if the body forces have a relatively large curl-free part (as for instance when hydrostatic forces are applied). Provided the observation time is sufficiently small so that $T \ll \mu^{-1}\ell_D^2$, a sharper velocity error estimate for the time-dependent Stokes equations can be derived by considering a different error decomposition. Taking inspiration from §53.3 and the above discussion on the error equation, we now consider the error decomposition resulting from the use of the projection operator $\boldsymbol{P}_h^{\mathrm{s}} : \boldsymbol{V} \to \boldsymbol{\mathcal{V}}_h$ defined in (73.11).

Theorem 73.6 (Velocity estimate). *Let $(\boldsymbol{u},p)$ solve (73.2). Assume that $\boldsymbol{u}\in H^1(J;\boldsymbol{V})$ and $p\in H^1(J;L^2_*(D))$. Let $(\boldsymbol{u}_h,p_h)$ solve (73.6). Set $\boldsymbol{e}_h(t):=\boldsymbol{u}_h(t)-\boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{u}(t))$ and $\boldsymbol{\eta}(t):=\boldsymbol{u}(t)-\boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{u}(t))$ for all $t\in J$. Assume that Q_h is H^1-conforming. The following holds for all $h\in\mathcal{H}$ with $c_5:=e(1+\sqrt{2})$,*

$$\|\boldsymbol{e}_h\|_{L^\infty(\overline{J};\boldsymbol{L}^2)}+2\sqrt{\mu}\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}\le c_5\|\boldsymbol{e}_h^0\|_{\boldsymbol{L}^2}\\ +(1+c_5)\sqrt{T}\big(\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}+\inf_{q_h\in H^1(J;Q_h)}\|\nabla(p-q_h)\|_{L^2(J;\boldsymbol{L}^2)}\big).\quad(73.16)$$

Proof. Let $q_h\in H^1(J;Q_h)$ and set $\zeta(t):=p(t)-q_h(t)$ for all $t\in J$. Using the test function $\boldsymbol{w}_h:=\boldsymbol{e}_h(t)$ for all $t\in J$ in (73.12) (notice that we have $\boldsymbol{w}_h\in\boldsymbol{\mathcal{V}}_h$) and using the coercivity of a gives

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2+2\mu\|\boldsymbol{e}_h\|_{\boldsymbol{V}}^2\le(\partial_t\boldsymbol{\eta},\boldsymbol{e}_h)_{\boldsymbol{L}^2}+b(\boldsymbol{e}_h,\zeta).$$

Using that Q_h is H^1-conforming, we have $b(\boldsymbol{e}_h,\zeta_h)=-(\nabla\zeta,\boldsymbol{e}_h)_{\boldsymbol{L}^2}$. Invoking the Cauchy–Schwarz and Young's inequality implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2+4\mu\|\boldsymbol{e}_h\|_{\boldsymbol{V}}^2\le T\big(\|\partial_t\boldsymbol{\eta}\|_{\boldsymbol{L}^2}^2+\|\nabla\zeta\|_{\boldsymbol{L}^2}^2\big)+\frac{2}{T}\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2}^2.$$

Using a simplified form of Gronwall's lemma (see Exercise 73.2) and taking the square root yields $\|\boldsymbol{e}_h\|_{L^\infty(\overline{J};\boldsymbol{L}^2)}\le e\big(\|\boldsymbol{e}_h^0\|_{\boldsymbol{L}^2}+\sqrt{T}(\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}+\|\nabla\zeta\|_{L^2(J;\boldsymbol{L}^2)})\big)$. Moreover, the above bound shows that $4\mu\|\boldsymbol{e}_h\|_{L^2(J;\boldsymbol{V})}^2\le T(\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2+\|\nabla\zeta\|_{L^2(J;\boldsymbol{L}^2)}^2)+2\|\boldsymbol{e}_h\|_{L^\infty(\overline{J};\boldsymbol{L}^2)}^2$. Taking the square root, putting everything together, and taking the infimum over $q_h\in H^1(J;Q_h)$ yields the assertion. □

Remark 73.7 (Theorem 73.4 vs. Theorem 73.6). The velocity error estimate from Theorem 73.6 does not use the viscous dissipation to bound the error, so that the estimate on $\|\boldsymbol{e}_h\|_{L^\infty(\overline{J};\boldsymbol{L}^2)}$ depends on the pressure approximation but not on the viscosity. More precisely, the pressure contribution in this estimate scales like $Th^{k'}|p|_{L^\infty(J;H^{k'+1})}$, where $k'\ge1$ is the degree of the finite elements used to approximate the pressure (see Remark 73.5). On the other hand the pressure contribution in the estimates of Theorem 73.4 depends on μ^{-1} since $\boldsymbol{\eta}$ therein depends on the pressure. More precisely, in the third estimate on $\|\boldsymbol{e}_h\|_{L^\infty(\overline{J};\boldsymbol{L}^2)}$, the pressure contribution scales like $h^s\ell_D^{1-s}\mu^{-1}h^{k'+1}|p|_{L^\infty(J;H^{k'+1})}$. Assuming for simplicity $s:=1$, one sees that the first estimate is smaller than the second if $T\mu\le h^2$. The practical consequence of this observation is that when the pressure term is dominant, one can observe the pre-asymptotic convergence rate $Th^{k'}$ when $\sqrt{T\mu}=:h_0\le h$, whereas the asymptotic rates predicted by Theorem 73.4 are recovered when the mesh is fine enough, i.e., $h\le h_0$. We refer the reader to Linke and Rebholz [216] for a discussion on this topic together with numerical experiments. □

73.3 Implicit Euler approximation

We show in this section how the semi-discrete problem (73.6) can be discretized in time by means of the implicit Euler scheme.

73.3.1 Discrete formulation

As in §73.2.1, the discretization in space uses the discrete spaces $(\boldsymbol{V}_h)_{h\in\mathcal{H}}$ and $(Q_h)_{h\in\mathcal{H}}$ built using the shape-regular sequence of matching meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$. The approximation setting is

conforming, i.e., $\boldsymbol{V}_h \subset \boldsymbol{V} := \boldsymbol{H}^1_0(D)$ and $Q_h \subset Q := L^2_*(D)$ for all $h \in \mathcal{H}$, and we assume that the inf-sup condition (73.4) holds true for all $h \in \mathcal{H}$. We use the notation introduced in §67.1.1 for the time discretization. We divide the time interval $J := (0,T)$, $T > 0$, into N subintervals J_n for all $n \in \mathcal{N}_\tau := \{1\!:\!N\}$. We take all these intervals to be of equal length for simplicity (although this is not a theoretical requirement). Thus, we define the time step as $\tau := \frac{T}{N}$, the discrete time nodes $t_n := n\tau$, for all $n \in \overline{\mathcal{N}}_\tau := \{0\!:\!N\}$, and we set $J_n := (t_{n-1}, t_n]$ for all $n \in \mathcal{N}_\tau$.

We assume $\boldsymbol{f} \in C^0(\overline{J}; \boldsymbol{L}^2(D))$ and we set $\boldsymbol{f}^n := \boldsymbol{f}(t_n) \in \boldsymbol{L}^2(D)$ for all $n \in \mathcal{N}_\tau$. We construct an approximating sequence $(\boldsymbol{u}_{h\tau}, p_{h\tau}) := (\boldsymbol{u}_h^n, p_h^n)_{n\in\mathcal{N}_\tau} \in (\boldsymbol{V}_h \times Q_h)^N$ as follows: First we set $\boldsymbol{u}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}_0, 0) = \boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{u}_0)$, then we compute $(\boldsymbol{u}_h^n, p_h^n) \in \boldsymbol{V}_h \times Q_h$ for all $n \in \mathcal{N}_\tau$ so that the following holds true:

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\boldsymbol{u}_h^n, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h^n) = (\boldsymbol{f}^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2}, \\ b(\boldsymbol{u}_h^n, q_h) = 0, \end{cases} \tag{73.17}$$

for all $(\boldsymbol{w}_h, q_h) \in \boldsymbol{V}_h \times Q_h$. Notice that $\boldsymbol{u}_h^n \in \boldsymbol{\mathcal{V}}_h$. At each time step, we must solve a problem of the following form:

$$\begin{cases} \widetilde{a}(\boldsymbol{u}_h^n, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, \tau\, p_h^n) = \boldsymbol{g}^n(\boldsymbol{w}_h), & \forall \boldsymbol{w}_h \in \boldsymbol{V}_h, \\ b(\boldsymbol{u}_h^n, q_h) = 0, & \forall q_h \in Q_h, \end{cases} \tag{73.18}$$

where $\widetilde{a}(\boldsymbol{u}_h^n, \boldsymbol{w}_h) := (\boldsymbol{u}_h^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + \tau a(\boldsymbol{u}_h^n, \boldsymbol{w}_h)$ and $\boldsymbol{g}^n(\boldsymbol{w}_h) := (\boldsymbol{u}_h^{n-1} + \tau \boldsymbol{f}^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2}$, i.e., at each time step we need to solve a time-independent Stokes-like problem similar to that described in Chapter 53. Since solving this saddle point problem at each time step may be computationally expensive, the reader is referred to Chapters 74 and 75 for more computationally effective techniques where the velocity and the pressure are uncoupled at each time step.

73.3.2 Algebraic realization and preconditioning

Let $\{\boldsymbol{\varphi}_i\}_{i\in\{1:I\}}$ be a basis of $\boldsymbol{V}_h$ with $I := \dim(\boldsymbol{V}_h)$. Let $\{\psi_k\}_{k\in\{1:K\}}$ be a basis of Q_h with $K := \dim(Q_h)$. Let $\mathsf{U}^n \in \mathbb{R}^I$ be the coordinate vector of $\boldsymbol{u}_h^n$ in the basis $\{\boldsymbol{\varphi}_i\}_{i\in\{1:I\}}$ for all $n \in \overline{\mathcal{N}}_\tau$, i.e., $\boldsymbol{u}_h^n(\boldsymbol{x}) := \sum_{i\in\{1:I\}} \mathsf{U}_i^n \boldsymbol{\varphi}_i(\boldsymbol{x})$. Let $\mathsf{P}^n \in \mathbb{R}^K$ be the coordinate vector of τp_h^n in the basis $\{\psi_k\}_{k\in\{1:K\}}$ for all $n \in \mathcal{N}_\tau$, i.e., $\tau p_h^n(\boldsymbol{x}) := \sum_{k\in\{1:K\}} P_k^n \psi_k(\boldsymbol{x})$. We introduce the stiffness matrix $\mathcal{A} \in \mathbb{R}^{I\times I}$ with $\mathcal{A}_{ij} := a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)$, the velocity mass matrix $\mathcal{M} \in \mathbb{R}^{I\times I}$ with $\mathcal{M}_{ij} := (\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)_{\boldsymbol{L}^2(D)}$, the divergence matrix $\mathcal{B} \in \mathbb{R}^{K\times I}$ with $\mathcal{B}_{kj} := b(\varphi_j, \psi_k)$, and the pressure mass matrix $\mathcal{N} \in \mathbb{R}^{K\times K}$ with $\mathcal{N}_{kl} = (\psi_l, \psi_k)_{L^2(D)}$, where $i, j \in \{1\!:\!I\}$ and $k, l \in \{1\!:\!K\}$. At each time step, the problem (73.18) is equivalent to solving the following linear system:

$$\begin{pmatrix} \mathcal{M} + \tau\mathcal{A} & \mathcal{B}^{\mathsf{T}} \\ \mathcal{B} & \mathbb{O}_{K,K} \end{pmatrix} \begin{pmatrix} \mathsf{U}^n \\ \mathsf{P}^n \end{pmatrix} = \begin{pmatrix} \mathsf{G}^n \\ 0 \end{pmatrix}, \tag{73.19}$$

where $\mathsf{G}^n := (\boldsymbol{g}^n(\varphi_i))_{i\in\{1:I\}} \in \mathbb{R}^I$ and $\mathbb{O}_{K,K}$ is the zero matrix in $\mathbb{R}^{K\times K}$.

It is shown in §50.2.2 that the above linear system amounts to solving

$$\mathcal{S}\mathsf{P}^n = \mathcal{B}(\mathcal{M} + \tau\mathcal{A})^{-1}\mathsf{G}^n, \tag{73.20}$$

with the Schur complement matrix $\mathcal{S} := \mathcal{B}(\mathcal{M} + \tau\mathcal{A})^{-1}\mathcal{B}^{\mathsf{T}}$. In Proposition 50.14, it is established that the condition number of $\mathcal{S}$, say $\kappa(\mathcal{S})$, is bounded from above by $\kappa(\mathcal{N})\frac{\|\tilde{a}\|\|b\|^2}{\widetilde{\alpha}\beta^2}$, where $\kappa(\mathcal{N})$ is the condition number of the pressure mass matrix, $\widetilde{\alpha}$ the coercivity constant of the modified bilinear form $\tilde{a}$ and $\|\widetilde{a}\|$ its boundedness constant, β the constant in the inf-sup condition (73.4),

and $\|b\|$ the boundedness constant of the bilinear form b. At this point we are facing a major difficulty. The only reasonable bound from below we can deduce on $\widetilde{\alpha}$ is $\widetilde{\alpha} \geq 2\tau\mu$. Moreover, $\|\tilde{a}\| \leq (\rho\mu + 2\tau\mu)$ where $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$. Thus, assuming that $\frac{\|b\|}{\beta} \sim 1$ and $\kappa(\mathcal{N}) \sim 1$, the bound from above on $\kappa(\mathcal{S})$ behaves like $\frac{\rho}{\tau}$, i.e., we must expect that this number grows unboundedly like $\frac{\rho}{\tau}$ when $\frac{\tau}{\rho} \to 0$. Actually, the definition of $\mathcal{S}$ shows that $\mathcal{S} \approx \mathcal{B}\mathcal{M}^{-1}\mathcal{B}^{\mathsf{T}}$ when $\frac{\tau}{\rho} \ll 1$, and one can then prove that $\kappa(\mathcal{S}) \sim h^{-2}\ell_D^2$, i.e., the condition number of $\mathcal{S}$ behaves like that of the Laplace operator. As a result, solving (73.20) by means of a standard gradient-based iterative technique entails very poor convergence rates as $\frac{\tau}{\rho} \to 0$; see Proposition 28.21. The situation worsens when the viscosity is so small that $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu} \gg T$, since in this case the time scale ratio $\frac{\rho}{T}$ becomes large and as a result $\frac{\tau}{\rho} = \frac{\tau}{T}\frac{T}{\rho}$ goes to zero even faster.

Two strategies are usually adopted in the literature to tackle this difficulty: either one preconditions (73.20) or one reformulates the time-stepping algorithm so as to uncouple the velocity and the pressure. We refer the reader to §50.2.3 and §50.3 for a brief overview of preconditioning techniques. Some uncoupling techniques are reviewed in Chapter 74.

73.3.3 Error analysis

We are going to proceed as in §67.1 for the error analysis of (73.17). For any Hilbert space B related to the velocity or the pressure, we consider the time-discrete norm $\|\phi_\tau\|_{\ell^2(J;B)}^2 := \sum_{n\in\mathcal{N}_\tau}\tau\|\phi^n\|_B^2$ with $\phi_\tau := (\phi^n)_{n\in\mathcal{N}_\tau} \in B^N$ (see (67.1)). Recall the time scale $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$.

Lemma 73.8 (Stability). *Let $(\boldsymbol{u}_{h\tau}, p_{h\tau}) \in (\boldsymbol{V}_h\times Q_h)^N$ solve (73.17) with $\boldsymbol{f}_\tau := (\boldsymbol{f}^n)_{n\in\mathcal{N}_\tau}$. The following stability estimates hold true:*

$$2\mu\|\boldsymbol{u}_{h\tau}\|_{\ell^2(J;\boldsymbol{V})}^2 + \|\boldsymbol{u}_h^N\|_{\boldsymbol{L}^2}^2 \leq \frac{\rho}{2}\|\boldsymbol{f}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2 + \|\boldsymbol{u}_h^0\|_{\boldsymbol{L}^2}^2, \tag{73.21a}$$

$$\|\boldsymbol{u}(t_n) - \boldsymbol{u}_h^n\|_{\boldsymbol{L}^2}^2 \leq e^{-\frac{t_n}{\rho}}\|\boldsymbol{u}_h^0\|_{\boldsymbol{L}^2}^2 + \frac{\rho}{2}\sum_{k\in\{1:n\}}\tau e^{-\frac{t_n-t_{k-1}}{\rho}}\|\boldsymbol{f}^k\|_{\boldsymbol{L}^2}^2, \tag{73.21b}$$

$$\|p_{h\tau}\|_{\ell^2(J;L^2)}^2 \leq \beta^{-2}\rho\mu\big(10\|\boldsymbol{f}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2 + 12\mu\|\boldsymbol{u}_0\|_{\boldsymbol{V}}^2\big). \tag{73.21c}$$

Proof. The proof of (73.21a) is the same as that of (67.7) (with $\alpha := 2\mu$). Indeed, taking the test function $\boldsymbol{w}_h := \boldsymbol{u}_h^n$ in (73.17) for all $n \in \mathcal{N}_\tau$, using that $b(\boldsymbol{u}_h^n, p_h^n) = 0$, and proceeding as in the parabolic setting leads to

$$\frac{1}{2}\|\boldsymbol{u}_h^n\|_{\boldsymbol{L}^2}^2 - \frac{1}{2}\|\boldsymbol{u}_h^{n-1}\|_{\boldsymbol{L}^2}^2 + \frac{1}{2}\|\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1}\|_{\boldsymbol{L}^2}^2 + 2\mu\tau\|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}^2 \leq \tau(\boldsymbol{f}^n, \boldsymbol{u}_h^n)_{\boldsymbol{L}^2}.$$

Then (73.21a) follows by using Young's inequality, summing over $n \in \mathcal{N}_\tau$, and using the telescopic form of the sum. The proof of (73.21b) is the same as that of (67.11). To derive the estimate on the pressure, we first obtain an estimate on the discrete time derivative of the velocity. Let us set $\delta_\tau\boldsymbol{u}_{h\tau} \in (\boldsymbol{V}_h)^N$ with $(\delta_\tau\boldsymbol{u}_{h\tau})^n := \frac{1}{\tau}(\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})$ for all $n \in \mathcal{N}_\tau$. We take the test function $\boldsymbol{v}_h := (\delta_\tau\boldsymbol{u}_{h\tau})^n$ in (73.17) and obtain after invoking the usual arguments

$$\|\delta_\tau\boldsymbol{u}_{h\tau}\|_{\ell^2(J;\boldsymbol{L}^2)}^2 \leq \|\boldsymbol{f}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2 + 2\mu\|\boldsymbol{u}_h^0\|_{\boldsymbol{V}}^2.$$

(Notice that the proof of (73.21a) already yields the much weaker bound $\tau\|\delta_\tau\boldsymbol{u}_{h\tau}\|_{\ell^2(J;\boldsymbol{L}^2)}^2 \leq \frac{\rho}{2}\|\boldsymbol{f}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2 + \|\boldsymbol{u}_h^0\|_{\boldsymbol{L}^2}^2$.) Then using the inf-sup condition (73.4), we infer that for all $n \in \mathcal{N}_\tau$,

$$\beta\|p_h^n\|_{L^2} \leq \frac{\ell_D}{C_{\mathrm{KPS}}}\big(\|(\delta_\tau\boldsymbol{u}_{h\tau})^n\|_{\boldsymbol{L}^2} + \|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}\big) + 2\mu\|\boldsymbol{u}_h^n\|_{\boldsymbol{V}}.$$

The bound on the pressure follows readily by proceeding as above. □

We are now in a position to establish error estimates.

Theorem 73.9 (Error estimates). *Let $(\boldsymbol{u},p)$ solve (73.2) and assume that $\boldsymbol{u}\in H^2(J;\boldsymbol{L}^2(D))\cap H^1(J;\boldsymbol{V})$ and $p\in H^1(J;L^2(D))$. Set $\boldsymbol{u}_\tau:=(\boldsymbol{u}(t_n))_{n\in\mathcal{N}_\tau}$ and $p_\tau:=(p(t_n))_{n\in\mathcal{N}_\tau}$. Let $(\boldsymbol{u}_{h\tau},p_{h\tau})\in(\boldsymbol{V}_h\times Q_h)^N$ solve (73.17). (i) There is c such that for all $h\in\mathcal{H}$, all $\tau>0$, and all $\mu>0$,*

$$\|\boldsymbol{u}_\tau-\boldsymbol{u}_{h\tau}\|_{\ell^2(J;\boldsymbol{V})}\le\|\boldsymbol{\eta}_\tau\|_{\ell^2(J;\boldsymbol{V})}+\frac{1}{\sqrt{2\mu}}\|\boldsymbol{\eta}_0\|_{\boldsymbol{L}^2}+\frac{\sqrt{2\rho}}{\sqrt{\mu}}\big(\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}+\tau\|\partial_{tt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}\big),\tag{73.22a}$$

$$\|p_\tau-p_{h\tau}\|_{\ell^2(J;L^2)}\le\|\zeta_\tau\|_{\ell^2(J;L^2)}+\frac{\sqrt{\rho\mu}}{\beta}\big(\sqrt{20}\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}+\sqrt{12\mu}\|\boldsymbol{\eta}_0\|_{\boldsymbol{V}}+\sqrt{20}\|\tau\|\partial_{tt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}\big).\tag{73.22b}$$

(ii) With $\boldsymbol{\eta}(t):=\boldsymbol{u}(t)-\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}(t),p(t))$, $\zeta(t):=p(t)-\mathcal{S}_h^{\mathrm{p}}(\boldsymbol{u}(t),p(t))$ for all $t\in J$, the Stokes elliptic projections $(\boldsymbol{\mathcal{S}}_h^{\mathrm{v}},\mathcal{S}_h^{\mathrm{p}})$ defined in (73.9), $\boldsymbol{\eta}_\tau:=(\boldsymbol{\eta}(t_n))_{n\in\mathcal{N}_\tau}$, $\zeta_\tau:=(\zeta(t_n))_{n\in\mathcal{N}_\tau}$, and $\boldsymbol{e}_h^0:=\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{0},p(0))$, the following holds true for all $n\in\mathcal{N}_\tau$:

$$\|\boldsymbol{u}_h^n\|_{\boldsymbol{L}^2}\le\|\boldsymbol{\eta}(t_n)\|_{\boldsymbol{L}^2}+e^{-\frac{t_n}{2\rho}}\|\boldsymbol{e}_h^0\|_{\boldsymbol{L}^2}+\sqrt{\rho}\big(\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}+\tau\|e^{-\frac{t_n-\cdot}{2\rho}}\partial_{tt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}\big).\tag{73.23}$$

Proof. Let us set $\boldsymbol{e}_h^n:=\boldsymbol{u}_h^n-\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}(t_n),p(t_n))$ and $\delta_h^n:=p_h^n-\mathcal{S}_h^{\mathrm{p}}(\boldsymbol{u}(t_n),p(t_n))$ for all $n\in\mathcal{N}_\tau$. Proceeding as in the proof of Theorem 67.6, we infer that

$$(\boldsymbol{e}_h^n-\boldsymbol{e}_h^{n-1},\boldsymbol{w}_h)_{\boldsymbol{L}^2}+\tau a(\boldsymbol{e}_h^n,\boldsymbol{w}_h)+\tau b(\boldsymbol{w}_h,\delta_h^n)=\tau(\boldsymbol{\xi}^n-\boldsymbol{\psi}^n,\boldsymbol{w}_h)_{\boldsymbol{L}^2},$$

with $\boldsymbol{\xi}^n:=\frac{1}{\tau}\int_{J_n}\partial_t\boldsymbol{\eta}(t)\mathrm{d}t$ and $\boldsymbol{\psi}^n:=-\frac{1}{\tau}\int_{J_n}(t-t_{n-1})\partial_{tt}\boldsymbol{u}(t)\mathrm{d}t$. Letting $\hat{\boldsymbol{f}}_\tau:=(\boldsymbol{\xi}^n-\boldsymbol{\psi}^n)_{n\in\mathcal{N}_\tau}$, we have (see again the proof of Theorem 67.6)

$$\|\hat{\boldsymbol{f}}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2\le2\|\partial_t\boldsymbol{\eta}\|_{L^2(J;\boldsymbol{L}^2)}^2+2\tau^2\|\partial_{tt}\boldsymbol{u}\|_{L^2(J;\boldsymbol{L}^2)}^2.$$

The error estimate (73.22a) follows by applying Lemma 73.8 and invoking the triangle inequality. Moreover, the proof of (73.23) is the same as for parabolic equations (see Theorem 67.9). Finally, we use the same arguments as in the proof of Theorem 73.8 to bound $\|\delta_{h\tau}\|_{\ell^2(J;L^2)}$ with $\delta_{h\tau}:=(\delta_h^n)_{n\in\mathcal{N}_\tau}$ and (73.22b) follows by invoking the triangle inequality. □

Remark 73.10 (Convergence rates). Under the assumptions and notation from Remark 73.5 the bounds from Theorem 73.9 imply that the error on the velocity in the $L^2(J;\boldsymbol{H}^1(D))$-norm and the error on the pressure in the $L^2(J;L^2(D))$-norm decay as $\mathcal{O}(h^r+\tau)$. Moreover, the error on the velocity in the $C^0(\overline{J};\boldsymbol{L}^2(D))$-norm decays as $\mathcal{O}(h^{r+s}+\tau)$, where $s\in(0,1]$ is the regularity pickup index. □

Remark 73.11 (Stabilization). All the stabilization techniques presented in Chapters 62 and 63 for the steady Stokes equations can be reused for the time-dependent (Navier–)Stokes equations. Examples include continuous interior penalty as in Burman and Fernández [65], local projection stabilization as in Arndt et al. [14], Dallmann et al. [99], Ahmed et al. [3, 4], and subgrid viscosity as in Guermond et al. [164]. Many other techniques can be used as well (see for instance Codina [90]), and the literature is prolific on the subject. We refer the reader to John [199, Chap. 8] for a review. □

73.4 Higher-order time approximation

We now briefly show how to achieve high-order accuracy in time by using the techniques developed in Chapters 69 and 70. To avoid duplicating the arguments for dG(k) schemes and cPG(k) schemes, and since we have shown in §69.2.4 and §70.1.3 that these methods are equivalent to implicit Runge–Kutta (IRK) techniques, we adopt the IRK point of view. We consider an s-stage IRK method defined by its Butcher coefficients $\{c_i\}_{i\in\{1:s\}}$, $\{b_i\}_{i\in\{1:s\}}$, $\{a_{ij}\}_{i,j\in\{1:s\}}$, and we set $t_{n,i} := t_{n-1} + c_i\tau$ for all $i \in \{1:s\}$ and all $n \in \mathcal{N}_\tau$, see (69.24) for Radau IIA IRK (i.e., dG(k) with $s := k+1$, $k \ge 0$) and (70.15) for KB IRK (i.e., cPG(k), with $s := k$, $k \ge 1$).

Our starting point is the constrained weak formulation (72.11). We do the approximation in space by using the setting described in §73.2.1. Let $\boldsymbol{A}_h^{\mathrm{St}} : \boldsymbol{\mathcal{V}}_h \to \boldsymbol{\mathcal{V}}_h$ be the operator s.t. $(\boldsymbol{A}_h^{\mathrm{St}}(\boldsymbol{v}_h), \boldsymbol{w}_h)_{\boldsymbol{L}^2} := a(\boldsymbol{v}_h, \boldsymbol{w}_h)$ for all $\boldsymbol{v}_h, \boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$ with $\boldsymbol{\mathcal{V}}_h$ defined in (73.5). We extend $\boldsymbol{A}_h^{\mathrm{St}}$ as an operator in $L^2(J;\boldsymbol{\mathcal{V}}_h)$ by setting $\boldsymbol{A}_h^{\mathrm{St}}(\boldsymbol{v}_h)(t) := \boldsymbol{A}_h^{\mathrm{St}}(\boldsymbol{v}_h(t))$ for all $t \in J$. We also define $\boldsymbol{f}_h^{\mathrm{St}} \in L^2(J;\boldsymbol{\mathcal{V}}_h)$ by $\int_J (\boldsymbol{f}_h^{\mathrm{St}}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}\mathrm{d}t := \int_J (\boldsymbol{f}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}\mathrm{d}t$ for all $\boldsymbol{w}_h \in \boldsymbol{\mathcal{V}}_h$. We then construct an IRK approximation of (73.6) as follows: First we set $\boldsymbol{u}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}_0, 0)$, then for all $n \in \mathcal{N}_\tau$, we solve the following set of coupled equations: Find $\{\boldsymbol{u}_h^{n,i}\}_{i\in\{1:s\}} \subset \boldsymbol{\mathcal{V}}_h$ s.t.

$$\boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n-1} = \tau \sum_{j\in\{1:s\}} a_{ij}\big(\boldsymbol{f}_h^{\mathrm{St}}(t_{n,j}) - \boldsymbol{A}_h^{\mathrm{St}}(\boldsymbol{u}_h^{n,j})\big). \tag{73.24}$$

Finally, $\boldsymbol{u}_h^n := \alpha_0\boldsymbol{u}_h^{n-1} + \sum_{i\in\{1:s\}} \alpha_i\boldsymbol{u}_h^{n,i}$, where $\alpha_i := \sum_{i\in\{1:s\}} b_j(a^{-1})_{ji}$ for all $i \in \{1:s\}$, $\alpha_0 := 1 - \sum_{i\in\{1:s\}} \alpha_i$, and $(a^{-1})_{ij}$ are the coefficients of the inverse of the Butcher matrix $(a_{ij})_{i,j\in\{1:s\}}$; see Remark 69.13.

Since constructing a basis for $\boldsymbol{\mathcal{V}}_h$ is in general difficult, let us reformulate the above technique using $\boldsymbol{V}_h$. We define $\boldsymbol{A}_h : \boldsymbol{V}_h \to \boldsymbol{V}_h$ and $B_h : \boldsymbol{V}_h \to Q_h$ by $(A_h(\boldsymbol{v}_h), \boldsymbol{w}_h)_{\boldsymbol{L}^2} := a(\boldsymbol{v}_h, \boldsymbol{w}_h)$ and $(B_h(\boldsymbol{v}_h), q_h)_{L^2} := b(\boldsymbol{v}_h, q_h)$ for all $\boldsymbol{v}_h, \boldsymbol{w}_h \in \boldsymbol{V}_h$ and all $q_h \in Q_h$. We finally define $\boldsymbol{f}_h \in L^2(J;\boldsymbol{V}_h)$ by $\int_J (\boldsymbol{f}_h, \boldsymbol{w}_h)_{\boldsymbol{L}^2}\mathrm{d}t = \int_J (\boldsymbol{f}, \boldsymbol{w}_h)_{\boldsymbol{L}^2}\mathrm{d}t$ for all $\boldsymbol{w}_h \in \boldsymbol{V}_h$. These definitions imply that

$$\sum_{j\in\{1:s\}} a_{ij}\big(\boldsymbol{A}_h^{\mathrm{St}}(\boldsymbol{u}_h^{n,j}) - \boldsymbol{A}_h(\boldsymbol{u}_h^{n,j}) + \boldsymbol{f}_h^{\mathrm{St}}(t_{n,j}) - \boldsymbol{f}_h(t_{n,j})\big) \in \ker(B_h)^\perp. \tag{73.25}$$

Since $\ker(B_h)^\perp = \mathrm{im}(B_h^*)$, we infer that (73.24) is equivalent to seeking pairs $\{(\boldsymbol{u}_h^{n,i}, p_h^{n,i})\}_{i\in\{1:s\}} \subset \boldsymbol{V}_h \times Q_h$ s.t.

$$\begin{cases} \boldsymbol{u}_h^{n,i} - \boldsymbol{u}_h^{n-1} = -\tau c_i B_h^*(p_h^{n,i}) + \tau \displaystyle\sum_{j\in\{1:s\}} a_{ij}\big(\boldsymbol{f}_h(t_{n,j}) - \boldsymbol{A}_h(\boldsymbol{u}_h^{n,j})\big), \\ B_h(\boldsymbol{u}_h^{n,i}) = 0. \end{cases} \tag{73.26}$$

Adopting the notation from §73.3.2, the algebraic realization of (73.26) using finite elements consists of solving for all $n \in \mathcal{N}_\tau$ the linear system

$$\left[\begin{array}{ccc|ccc} \mathcal{M}+\tau a_{11}\mathcal{A} & \cdots & \tau a_{1s}\mathcal{A} & \mathcal{B}^{\mathsf{T}} & \cdots & \mathbb{O}_{I\times K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \tau a_{s1}\mathcal{A} & \cdots & \mathcal{M}+\tau a_{ss}\mathcal{A} & \mathbb{O}_{I\times K} & \cdots & \mathcal{B}^{\mathsf{T}} \\ \hline \mathcal{B} & \cdots & \mathbb{O}_{K\times I} & \mathbb{O}_{K\times K} & \cdots & \mathbb{O}_{K\times K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{O}_{K\times I} & \cdots & \mathcal{B} & \mathbb{O}_{K\times K} & \cdots & \mathbb{O}_{K\times K} \end{array}\right] \left[\begin{array}{c} \mathsf{U}^{n,1} \\ \vdots \\ \mathsf{U}^{n,s} \\ \hline \mathsf{P}^{n,1} \\ \vdots \\ \mathsf{P}^{n,s} \end{array}\right] = \left[\begin{array}{c} \mathsf{G}^{n,1} \\ \vdots \\ \mathsf{G}^{n,s} \\ \hline 0 \\ \vdots \\ 0 \end{array}\right].$$

Here, $\mathsf{U}^{n,i} \in \mathbb{R}^I$ is the coordinate vector of $\boldsymbol{u}_h^{n,i}$ in the basis $\{\boldsymbol{\varphi}_i\}_{i\in\{1:I\}}$ and $\mathsf{P}^{n,i} \in \mathbb{R}^K$ is the coordinate vector of $\tau c_i p_h^{n,i}$ in the basis $\{\psi_k\}_{k\in\{1:K\}}$ for all $n \in \mathcal{N}_\tau$ and all $i \in \{1:s\}$. The entries

of the load vector are $\mathsf{G}_i^{n,l} := \mathcal{M}\mathsf{U}_i^{n-1} + \tau \sum_{m\in\{1:s\}} a_{lm}\mathsf{F}_i^{n,m}$ with $\mathsf{F}_i^{n,m} := (\boldsymbol{f}(t_{n,m}), \boldsymbol{\varphi}_i)_{\boldsymbol{L}^2}$ for all $i \in \{1{:}I\}$.

Exercises

Exercise 73.1 (Well-posedness). Prove Proposition 73.1. (*Hint*: adapt the proof of Theorem 72.3.)

Exercise 73.2 (Simplified Gronwall's lemma). Let $a \in W^{1,1}(J;\mathbb{R})$, let $b \in L^\infty(\overline{J};\mathbb{R})$, and let $\gamma > 0$. Assume that $\frac{\mathrm{d}}{\mathrm{d}t}a(t) \le \frac{1}{\gamma}a(t) + b(t)$ for all $t \in \overline{J}$. Prove that $a(t) \le e^{\frac{t}{\gamma}}\big(a(0) + \min(t,\gamma)\|b\|_{L^\infty(\overline{J}_t)}\big)$ with $\overline{J}_t := (0,t)$ for all $t \in \overline{J}$. (*Hint*: observe that $\int_0^t e^{\frac{t-s}{\gamma}}\,\mathrm{d}s \le \min(t,\gamma)e^{\frac{t}{\gamma}}$.) *Note*: this is a simplified form of Gronwall's lemma; see Exercise 65.3.

Exercise 73.3 (BDF2, Crank–Nicolson). (i) Using the setting described in §68.2 for BDF2, write the discrete formulation and the algebraic realization of the time-dependent Stokes equations with the time discretization performed with BDF2. (ii) Same question for the Crank–Nicolson scheme using the setting described in §68.3. (iii) Same question for the Crank–Nicolson scheme using the setting described in §73.4.

Chapter 74

Projection methods

The goal of this chapter and the next one is to give a brief overview of some splitting techniques to approximate the time-dependent Stokes problem in time. The common feature of the algorithms is that each time step leads to subproblems where the velocity and the pressure are uncoupled. The linear algebra resulting from the space approximation is therefore simplified, making these methods attractive for their efficiency. In this chapter, we review a class of techniques known in the literature as *projection methods* where the accuracy in time is limited to second order. The algorithms reviewed in the next chapter are based on an artificial compressibility perturbation of the mass conservation equation and can reach arbitrary accuracy in time. Projection methods are among the most popular strategies to discretize in time the time-dependent Stokes equations. These methods have been pioneered in the work of Chorin [83, 84] and Temam [271]. The material in this chapter is adapted from Guermond et al. [165].

74.1 Model problem and Helmholtz decomposition

For simplicity, we assume that homogeneous Dirichlet boundary conditions are enforced on the velocity over the entire boundary and that the viscosity μ is constant. We consider the mixed weak formulation (73.2), i.e., we assume that the source term satisfies $\boldsymbol{f} \in L^2(J;\boldsymbol{L}^2(D))$ and that the initial condition satisfies $\boldsymbol{u}_0 \in \boldsymbol{\mathcal{V}} := \{\boldsymbol{v} \in \boldsymbol{V} \mid \nabla\cdot\boldsymbol{v} = 0\}$ with $\boldsymbol{V} := \boldsymbol{H}_0^1(D)$. Denoting $(\boldsymbol{u}, p)$ the weak solution, we have $\boldsymbol{u} \in L^2(J;\boldsymbol{\mathcal{V}})$, $\partial_t\boldsymbol{u} \in L^2(J;\boldsymbol{L}^2(D))$, and $p \in L^2(J;Q)$ with $Q := L^2_*(D) := \{q \in L^2(D) \mid \int_D q\,\mathrm{d}x = 0\}$ (see Theorem 72.3). Recall that the spaces $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{V}$ are equipped with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$, and Q is equipped with the L^2-norm. Combining Korn's inequality with the Poincaré–Steklov inequality, we have $C_{\mathrm{KPS}}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \le \ell_D\|\boldsymbol{v}\|_{\boldsymbol{V}}$ for all $\boldsymbol{v} \in \boldsymbol{V}$, where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. We define the time scale $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$.

Let us state a decomposition of $\boldsymbol{L}^2(D)$ that plays an important role in projection methods. We define the following spaces:

$$H^1_*(D) := H^1(D) \cap L^2_*(D), \qquad \boldsymbol{\mathcal{H}} := \{\boldsymbol{v} \in \boldsymbol{L}^2(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D}\cdot\boldsymbol{n} = 0\},$$

where $\nabla\cdot\boldsymbol{v} = 0$ and $\boldsymbol{v}_{|\partial D}\cdot\boldsymbol{n} = 0$ mean that $(\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(D)} = 0$ for all $q \in H^1_*(D)$.

Lemma 74.1 (Helmholtz decomposition). *The following $\boldsymbol{L}^2$-orthogonal decomposition holds true:*

$$\boldsymbol{L}^2(D) = \boldsymbol{\mathcal{H}} \oplus \nabla(H^1_*(D)). \tag{74.1}$$

The $\boldsymbol{L}^2$-orthogonal projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}} : \boldsymbol{L}^2(D) \to \boldsymbol{\mathcal{H}}$ resulting from (74.1) *is often called* Leray projection *in the literature.*

Proof. Let $\boldsymbol{v} \in \boldsymbol{L}^2(D)$. To project $\boldsymbol{v}$ onto $\nabla(H^1_*(D))$, we pose the following problem: Find $p \in H^1_*(D)$ s.t. $(\nabla p, \nabla q)_{\boldsymbol{L}^2(D)} = (\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(D)}$ for all $q \in H^1_*(D)$. Then we set $\boldsymbol{u} := \boldsymbol{v} - \nabla p$. By construction, we have $\boldsymbol{u} \in \boldsymbol{\mathcal{H}}$ since $\boldsymbol{u} \in \boldsymbol{L}^2(D)$ and $(\boldsymbol{u}, \nabla q)_{\boldsymbol{L}^2(D)} = 0$ for all $q \in H^1_*(D)$. The orthogonality of the decomposition $\boldsymbol{v} = \boldsymbol{u} + \nabla p$ follows from the construction. □

74.2 Pressure correction in standard form

We use the same notation as in §73.3.1 to describe the time discretization, and for the time being the space variable is not discretized. Recall that the time interval $J := (0, T)$, $T > 0$, is divided into N subintervals J_n for all $n \in \mathcal{N}_\tau := \{1\!:\!N\}$. We assume that the time step $\tau := \frac{T}{N}$ is constant. We set $t_n := n\tau$ for all $n \in \overline{\mathcal{N}}_\tau := \{0\!:\!N\}$, and $J_n := (t_{n-1}, t_n]$ for all $n \in \mathcal{N}_\tau$. We approximate the time derivative of the velocity with the Backward Difference Formula of order q (BDFq) as $\partial_t \boldsymbol{u}(t_n) = \frac{1}{\tau}(\beta_q \boldsymbol{u}(t_n) - \sum_{j\in\{1:q\}} \beta_{j-1} \boldsymbol{u}(t_{n-j})) + \mathcal{O}(\tau^q)$ with $q \in \{1, 2\}$. For $q := 1$, we set $\beta_1 := 1$, $\beta_0 := 1$ (i.e., BDF1 is the implicit Euler scheme), and for $q := 2$, we have $\beta_2 := \frac{3}{2}$, $\beta_1 := 2$, $\beta_0 := -\frac{1}{2}$ (i.e., BDF2 is the time-stepping scheme studied in §68.2 for parabolic equations). For simplicity, we assume that $\boldsymbol{f} \in C^0(\overline{J}; \boldsymbol{L}^2(D))$ and we set $\boldsymbol{f}^n := \boldsymbol{f}(t_n) \in \boldsymbol{L}^2(D)$ for all $n \in \mathcal{N}_\tau$.

74.2.1 Formulation of the method

Let $n \in \mathcal{N}_\tau$. In a projection method, each time step from t_{n-1} to t_n is composed of three substeps. In the first substep, the pressure is made explicit by using some *extrapolation formula*, and a provisional velocity field $\widetilde{\boldsymbol{u}}^n$ is computed using the momentum equation. The extrapolated pressure is denoted by $p^{\star,n}$, and the two most frequent choices are $p^{\star,n} := 0$ (zero-order extrapolation) and $p^{\star,n} := p^{n-1}$ (first-order extrapolation). In the second substep, the velocity field $\boldsymbol{u}^n$ is obtained by projecting the provisional velocity field $\widetilde{\boldsymbol{u}}^n$ onto the space of incompressible (divergence-free) vector fields by using the Leray projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}$. The pressure p^n is updated in the third substep.

The method, known in the literature as *pressure-correction method in standard form*, proceeds as follows. One sets $\boldsymbol{u}^0 := \boldsymbol{u}_0$ and if first-order pressure extrapolation is used, one assumes that $p(0)$ is available (see Remark 74.4) and one sets $p^{*,0} := p(0)$. Then one generates the three sequences $\widetilde{\boldsymbol{u}}_\tau := (\widetilde{\boldsymbol{u}}^n)_{n\in\mathcal{N}_\tau} \in (\boldsymbol{V})^N$, $\boldsymbol{u}_\tau := (\boldsymbol{u}^n)_{n\in\mathcal{N}_\tau} \in (\boldsymbol{\mathcal{H}})^N$, $p_\tau := (p^n)_{n\in\mathcal{N}_\tau} \in (Q)^N$ by performing for all $n \in \mathcal{N}_\tau$ the following three substeps:

1. One computes $\widetilde{\boldsymbol{u}}^n \in \boldsymbol{V} := \boldsymbol{H}^1_0(D)$ such that

$$\frac{1}{\tau}\Big(\beta_q \widetilde{\boldsymbol{u}}^n - \sum_{j\in\{1:q\}} \beta_{j-1} \boldsymbol{u}^{n-j}\Big) - \nabla\cdot\mathbb{s}(\widetilde{\boldsymbol{u}}^n) + \nabla p^{\star,n} = \boldsymbol{f}^n. \tag{74.2}$$

   If one uses BDF2, then one sets $q := 2$ if $n \in \mathcal{N}_\tau$, $n \geq 2$, and $q := 1$ if $n = 1$, whereas if one uses BDF1, one sets $q := 1$ for all $n \in \mathcal{N}_\tau$. The weak form of (74.2) is

$$\beta_q(\widetilde{\boldsymbol{u}}^n, \boldsymbol{w})_{\boldsymbol{L}^2} + \tau a(\widetilde{\boldsymbol{u}}^n, \boldsymbol{w}) + b(\boldsymbol{w}, \tau p^{\star,n}) = (\boldsymbol{g}^n, \boldsymbol{w})_{\boldsymbol{L}^2}, \tag{74.3}$$

   for all $\boldsymbol{w} \in \boldsymbol{V}$ with $\boldsymbol{g}^n := \tau \boldsymbol{f}^n + \sum_{j\in\{1:q\}} \beta_{j-1} \boldsymbol{u}^{n-j}$ and the bilinear forms $a(\boldsymbol{v}, \boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)}$ and $b(\boldsymbol{v}, q) := -(q, \nabla\cdot\boldsymbol{v})_{L^2(D)}$.

2. Since the divergence of $\widetilde{\boldsymbol{u}}^n$ is (in general) nonzero, this vector field is then projected onto $\boldsymbol{\mathcal{H}}$, i.e., we set $\boldsymbol{u}^n := \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\widetilde{\boldsymbol{u}}^n)$. Owing to Lemma 74.1, the second substep thus consists of seeking $\boldsymbol{u}^n \in \boldsymbol{\mathcal{H}}$ and $\phi^n \in H^1_*(D)$ s.t.

$$\boldsymbol{u}^n + \tau\nabla\phi^n = \widetilde{\boldsymbol{u}}^n, \quad \nabla\cdot\boldsymbol{u}^n = 0, \quad \boldsymbol{u}^n_{|\partial D}\cdot\boldsymbol{n} = 0. \tag{74.4}$$

   In practice, one first solves a Poisson equation with Neumann conditions, i.e., one seeks $\phi^n \in H^1_*(D)$ s.t. $\tau(\nabla\phi^n, \nabla q)_{\boldsymbol{L}^2(D)} = (\widetilde{\boldsymbol{u}}^n, \nabla q)_{\boldsymbol{L}^2(D)}$ for all $q \in H^1_*(D)$, and then one sets $\boldsymbol{u}^n := \widetilde{\boldsymbol{u}}^n - \tau\nabla\phi^n$.

3. The pressure is updated by setting

$$p^n := \beta_q\phi^n + p^{\star,n}. \tag{74.5}$$

It is in general more computationally effective to solve the two independent problems (74.2) and (74.4) than to solve the coupled problem (73.17) resulting from the use of the implicit Euler method for the fully coupled problem. Indeed, (74.2) corresponds to one implicit step for a parabolic problem and (74.4) just amounts to solving an elliptic problem, whereas (73.17) requires solving a saddle point problem.

To motivate (74.5), we multiply the first equation in (74.4) by $\beta_q\frac{1}{\tau}$ and add the result to (74.2). This yields

$$D_\tau^{(q)}\boldsymbol{u}^n - \nabla\cdot\mathbb{s}(\widetilde{\boldsymbol{u}}^n) + \nabla\left(\beta_q\phi^n + p^{\star,n}\right) = \boldsymbol{f}^n, \tag{74.6}$$

with $D_\tau^{(q)}\boldsymbol{u}^n := \frac{1}{\tau}(\beta_q\boldsymbol{u}^n - \sum_{j\in\{1:q\}}\beta_{j-1}\boldsymbol{u}^{n-j})$, i.e., $D_\tau^{(1)}\boldsymbol{u}^n := \frac{1}{\tau}(\boldsymbol{u}^n - \boldsymbol{u}^{n-1})$ and $D_\tau^{(2)}\boldsymbol{u}^n := \frac{1}{\tau}(\frac{3}{2}\boldsymbol{u}^n - 2\boldsymbol{u}^{n-1} + \frac{1}{2}\boldsymbol{u}^{n-2})$. Using (74.5) in (74.6) leads to the following consistent approximation of the momentum conservation equation: $D_\tau^{(q)}\boldsymbol{u}^n - \nabla\cdot\mathbb{s}(\widetilde{\boldsymbol{u}}^n) + \nabla p^n = \boldsymbol{f}^n$.

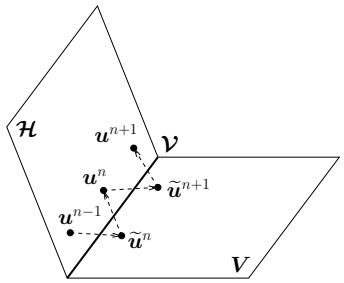

Figure 74.1: Projection algorithm. At step n, ones computes $\widetilde{\boldsymbol{u}}^n \in \boldsymbol{V}$ with full Dirichlet boundary conditions, but without enforcing incompressibility so that $\widetilde{\boldsymbol{u}}^n$ pops out of $\boldsymbol{\mathcal{H}}$. Then one projects $\widetilde{\boldsymbol{u}}^{n+1}$ onto $\boldsymbol{\mathcal{H}}$ to enforce incompressibility and the Dirichlet condition on the normal component, but the tangential component of $\boldsymbol{u}^n$ may be nonzero so that $\boldsymbol{u}^n$ pops out of $\boldsymbol{V}$. Neither $\widetilde{\boldsymbol{u}}^n$ nor $\boldsymbol{u}^n$ is in general in $\boldsymbol{\mathcal{V}} = \boldsymbol{V}\cap\boldsymbol{\mathcal{H}}$.

Remark 74.2 ($\widetilde{\boldsymbol{u}}^n$ vs. $\boldsymbol{u}^n$). The velocity $\widetilde{\boldsymbol{u}}^n \in \boldsymbol{V}$ is an approximation of $\boldsymbol{u}(t_n)$ that satisfies the boundary conditions but is not divergence-free. This defect is corrected by projecting $\widetilde{\boldsymbol{u}}^n$ onto $\boldsymbol{\mathcal{H}}$ (whence the name of the method). Although $\boldsymbol{u}^n$ is divergence-free, it is not necessarily a better approximation of $\boldsymbol{u}(t_n)$ since it does not satisfy the no-slip boundary condition, i.e., its tangential component is in general nonzero. Moreover, $\boldsymbol{u}^n$ is not necessarily in $\boldsymbol{V}$. Hence, neither $\widetilde{\boldsymbol{u}}^n$ nor $\boldsymbol{u}^n$ is in general in $\boldsymbol{\mathcal{V}} = \boldsymbol{V}\cap\boldsymbol{\mathcal{H}}$. A schematic representation of the pressure-correction algorithm (or projection method) is shown in Figure 74.1. □

Remark 74.3 (Elimination of $\boldsymbol{u}^n$). It is possible to avoid computing the sequence $\boldsymbol{u}_\tau$ by using $\boldsymbol{u}^{n-j} = \widetilde{\boldsymbol{u}}^{n-j} - \tau\nabla\phi^{n-j}$ for all $j \in \{1\!:\!q\}$. Then for all $n \geq q$, one rewrites (74.2) as follows:

$$D_\tau^{(q)}\widetilde{\boldsymbol{u}}^n - \nabla\cdot\mathbb{s}(\widetilde{\boldsymbol{u}}^n) + \nabla\Big(p^{\star,n} + \sum_{j\in\{1\,:\,q\}}\beta_{j-1}\phi^{n-j}\Big) = \boldsymbol{f}^n. \tag{74.7}$$

The projection step (74.4) can be solved without invoking $\boldsymbol{u}^n$ as follows:

$$-\Delta\phi^n = -\tau^{-1}\nabla\cdot\widetilde{\boldsymbol{u}}^n, \quad \boldsymbol{n}\cdot\nabla\phi^n_{|\partial D} = 0. \tag{74.8}$$

The pressure update is unchanged: $p^n = \beta_q\phi^n + p^{\star,n}$. The scheme (74.7)-(74.8)-(74.5) is equivalent to (74.2)-(74.4)-(74.5), but it is somewhat easier to implement; see §74.4. □

Remark 74.4 (Initial pressure). The algorithm proposed by Chorin [83, 84] and Temam [271] uses the zero-order pressure extrapolation $p^{\star,n} := 0$ and BDF1. Theorem 74.7 shows that the accuracy is rather poor. The accuracy is improved by using the first-order pressure extrapolation $p^{\star,n} := p^{n-1}$ with BDF1, as pointed out in Goda [137]. Notice that algorithms based on first-order pressure extrapolation assume more smoothness than the *Chorin–Temam algorithm* since they require the existence of $p(0)$ in some reasonably smooth space, although $p(0)$ is not an initial data for the time-dependent Stokes problem (72.1). If $\boldsymbol{u}_0 \in \boldsymbol{H}^2(D)\cap\boldsymbol{\mathcal{H}}$, one can compute $p(0)$ by solving $(\nabla p(0), \nabla q)_{\boldsymbol{L}^2(D)} = (\boldsymbol{f}(0) + \nabla\cdot\mathbb{s}(\boldsymbol{u}_0), \nabla q)_{\boldsymbol{L}^2(D)}$ for all $q \in H^1_*(D)$. Notice that $(\nabla\cdot\mathbb{s}(\boldsymbol{u}_0), \nabla q)_{\boldsymbol{L}^2(D)}$ is in general nonzero; indeed the field $\nabla\cdot\mathbb{s}(\boldsymbol{u}_0)$ is divergence-free but its normal component at ∂D is in general nonzero. □

74.2.2 Stability and convergence properties

Lemma 74.5 (Stability). *Let $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ solve* (74.2)-(74.4)-(74.5) *with* BDFq, $q \in \{1, 2\}$, *and the first-order pressure extrapolation $p^{\star,n} := p^{n-1}$. Let $\boldsymbol{f}_\tau := (\boldsymbol{f}^n)_{n\in\mathcal{N}_\tau}$. There is c such that for all $\tau > 0$ ($c = 1$ for* BDF1*),*

$$\|\boldsymbol{u}^N\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla p^N\|^2_{\boldsymbol{L}^2} + 2\mu\|\widetilde{\boldsymbol{u}}_\tau\|^2_{\ell^2(J;\boldsymbol{V})} \leq c\left(\frac{\rho}{2}\|\boldsymbol{f}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)} + \|\boldsymbol{u}^0\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla p^0\|^2_{\boldsymbol{L}^2}\right). \tag{74.9}$$

Proof. We restrict ourselves to BDF1 for brevity. The proof for BDF2 is similar. Testing (74.2) with $2\tau\widetilde{\boldsymbol{u}}^n$, using the coercivity of the bilinear form $a(\boldsymbol{v}, \boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)}$ on $\boldsymbol{V}$, and the algebraic identity (67.9) already invoked in the context of the implicit Euler scheme, we obtain

$$\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2} - \|\boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2} + 4\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \leq 2\tau(\boldsymbol{f}^n, \widetilde{\boldsymbol{u}}^n)_{\boldsymbol{L}^2} + 2\tau(p^{n-1}, \nabla\cdot\widetilde{\boldsymbol{u}}^n)_{L^2}.$$

Since

$$2\tau(\boldsymbol{f}^n, \widetilde{\boldsymbol{u}}^n)_{\boldsymbol{L}^2} \leq \frac{\tau}{2\mu}\|\boldsymbol{f}^n\|^2_{\boldsymbol{V}'} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \leq \frac{\tau\rho}{2}\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}},$$

where we used Young's inequality, the bound $\|\boldsymbol{f}^n\|_{\boldsymbol{V}'} \leq C^{-1}_{\mathrm{KPS}}\ell_D\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}$, and the definition of the time scale ρ, we infer that

$$\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2} - \|\boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2} + 2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{V}} \leq \frac{\tau\rho}{2}\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2} + 2\tau(p^{n-1}, \nabla\cdot\widetilde{\boldsymbol{u}}^n)_{L^2}.$$

Using that $\phi^n = p^n - p^{n-1}$ since $\beta_q = \beta_1 := 1$ for BDF1, we recast (74.4) as $\boldsymbol{u}^n + \tau\nabla p^n = \widetilde{\boldsymbol{u}}^n + \tau\nabla p^{n-1}$. We square this identity, integrate over D, and use that $\boldsymbol{u}^n$ is divergence-free to obtain

$$\|\boldsymbol{u}^n\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla p^n\|^2_{\boldsymbol{L}^2} = \|\widetilde{\boldsymbol{u}}^n\|^2_{\boldsymbol{L}^2} - 2\tau(p^{n-1}, \nabla\cdot\widetilde{\boldsymbol{u}}^n)_{L^2} + \tau^2\|\nabla p^{n-1}\|^2_{\boldsymbol{L}^2}.$$

Summing this identity and the above estimate yields

$$\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2+\tau^2\|\nabla p^n\|_{\boldsymbol{L}^2}^2+2\mu\tau\|\widetilde{\boldsymbol{u}}^n\|_{\boldsymbol{V}}^2\le\frac{\tau\rho}{2}\|\boldsymbol{f}^n\|_{\boldsymbol{L}^2}^2+\|\boldsymbol{u}^{n-1}\|_{\boldsymbol{L}^2}^2+\tau^2\|\nabla p^{n-1}\|_{\boldsymbol{L}^2}^2.$$

Summing the result over $n\in\mathcal{N}_\tau$ yields the assertion. □

Remark 74.6 (Zero-order pressure extrapolation). Using BDF1 and the zero-order pressure extrapolation $p^{\star,n}:=0$, the stability estimate in Lemma 74.5 reduces to

$$\|\boldsymbol{u}^N\|_{\boldsymbol{L}^2}^2+\tau^2\|\nabla p_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2+2\mu\|\widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{V})}^2\le\frac{\rho}{2}\|\boldsymbol{f}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}^2+\|\boldsymbol{u}^0\|_{\boldsymbol{L}^2}^2.$$

The proof of this inequality is simpler than that of (74.9) since there is no need to bound $(p^{n-1},\nabla\cdot\widetilde{\boldsymbol{u}}^n)_{L^2}$; see Exercise 74.1. □

Let us now review some convergence results. We introduce the discrete time sequences $\boldsymbol{\pi}_\tau(\boldsymbol{u}):=(\boldsymbol{u}(t_n))_{n\in\mathcal{N}_\tau}$ and $\pi_\tau(p):=(p(t_n))_{n\in\mathcal{N}_\tau}$, where $(\boldsymbol{u},p)$ denotes the solution to (73.2). Moreover, $c(\boldsymbol{u},p,T)$ denotes a generic constant that depends on $\boldsymbol{u}$, p, and T, but is independent of τ.

Theorem 74.7 (Convergence: BDF1, $p^{\star,n}:=0$). *Assume that the solution $(\boldsymbol{u},p)$ to (73.2) is sufficiently smooth. Then the sequences $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ generated by (74.2)-(74.4)-(74.5), with* BDF1 *and the zero-order pressure extrapolation $p^{\star,n}:=0$, satisfy*

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\boldsymbol{u}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)}+\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)}\le c(\boldsymbol{u},p,T)\tau,$$
$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{V})}+\|\pi_\tau(p)-p_\tau\|_{\ell^2(J;Q)}\le c(\boldsymbol{u},p,T)\tau^{\frac12}.$$

Proof. See Rannacher [240]. (Note that in general $\boldsymbol{u}_\tau\notin(\boldsymbol{V})^N$.) □

Despite its simplicity, using the zero-order pressure extrapolation $p^{\star,n}:=0$ is not satisfactory since the convergence rate is limited to $\mathcal{O}(\tau)$ for the velocity in the $\boldsymbol{L}^2$-norm and $\mathcal{O}(\tau^{\frac12})$ for the velocity in the $\boldsymbol{H}^1$-norm and the pressure in the L^2-norm. The convergence loss comes from the fact that the method is basically a first-order artificial compressibility technique as shown in Rannacher [240], Shen [252]. Numerous variants have been proposed to cure this problem. One of them consists of using the first-order pressure extrapolation $p^{\star,n}:=p^{n-1}$.

Theorem 74.8 (Convergence: BDF1, $p^{\star,n}:=p^{n-1}$). *Assume that the solution $(\boldsymbol{u},p)$ to (73.2) is sufficiently smooth. Set $\boldsymbol{u}^0:=\boldsymbol{u}(0)=\boldsymbol{u}_0$ and $p^0:=p(0)$. Then the sequences $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ generated by (74.2)-(74.4)-(74.5), with* BDF1 *and the first-order pressure extrapolation $p^{\star,n}:=p^{n-1}$, satisfy*

$$\|\boldsymbol{\pi}_\tau(u)-\boldsymbol{u}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)}+\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)}\le c(\boldsymbol{u},p,T)\tau,$$
$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{V})}+\|\pi_\tau(p)-p_\tau\|_{\ell^2(J;Q)}\le c(\boldsymbol{u},p,T)\tau.$$

Proof. See Shen [253], Guermond and Quartapelle [160]. □

The first-order pressure extrapolation $p^{\star,n}:=p^{n-1}$ became popular after it was informally shown in van Kan [280] to increase the accuracy of the method when used together with a second-order time-stepping scheme.

Theorem 74.9 (Convergence: BDF2, $p^{\star,n}:=p^{n-1}$). *Assume that the solution $(\boldsymbol{u},p)$ to (73.2) is sufficiently smooth. Set $\boldsymbol{u}^0:=\boldsymbol{u}(0)=\boldsymbol{u}_0$ and $p^0:=p(0)$. Then the sequences $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ generated by (74.2)-(74.4)-(74.5), with* BDF2 *and the first-order pressure extrapolation $p^{\star,n}:=p^{n-1}$, satisfy*

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\boldsymbol{u}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}+\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{L}^2)}\le c(\boldsymbol{u},p,T)\tau^2,$$
$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{V})}+\|\pi_\tau(p)-p_\tau\|_{\ell^2(J;Q)}\le c(\boldsymbol{u},p,T)\tau.$$

Proof. See Shen [256], Guermond [145]; see also E and Liu [115], Strikwerda and Lee [266], Brown et al. [56] for different proofs based on normal mode analysis in the half-plane or in a periodic channel. □

Remark 74.10 (Higher-order extrapolations). Higher-order pressure extrapolations, like $p^{\star,n} := 2p^{n-1} - p^{n-2}$ for $n \geq 2$, have been considered in the literature. Whether using a second-order or higher-order pressure extrapolation yields a stable scheme is not yet clear. At the time of this writing no proof of stability has been published. A singular perturbation argument advanced in Shen [254] actually indicates that some of these higher-order extrapolation algorithms should not be stable for small time steps. This issue is an open question. □

74.3 Pressure correction in rotational form

Theorem 74.9 shows that the scheme (74.2)-(74.4)-(74.5) with BDF2 and the first-order pressure extrapolation is second-order accurate on the velocity in the $\boldsymbol{L}^2$-norm, but it is only first-order accurate in the $\boldsymbol{H}^1$-norm. The reason for this loss of convergence is a numerical boundary layer effect. Actually, we observe from (74.4)-(74.5) that $\boldsymbol{n}\cdot\nabla(p^n - p^{*,n})_{|\partial D} = 0$. If the first-order pressure extrapolation is used, this implies that

$$\boldsymbol{n}\cdot\nabla p^n_{|\partial D} = \boldsymbol{n}\cdot\nabla p^{n-1}_{|\partial D} = \cdots = \boldsymbol{n}\cdot\nabla p^0_{|\partial D}. \tag{74.10}$$

It is this nonrealistic Neumann boundary condition on the pressure that introduces the boundary layer in question and consequently limits the accuracy of the scheme.

74.3.1 Formulation of the method

Projection methods in rotational form exploit that the viscosity μ is constant (as we are assuming here). Using that $\nabla\cdot\boldsymbol{u} = 0$, one observes that $\nabla\cdot\mathbb{s}(\boldsymbol{u}) = \mu\nabla\cdot(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathsf{T}}) = \mu\Delta\boldsymbol{u}$ since $\boldsymbol{u}$ is divergence-free. One key ingredient to derive a more accurate algorithm is to use the rotational form of the vector Laplacian in (74.2), namely, $-\Delta\widetilde{\boldsymbol{u}}^n = -\nabla(\nabla\cdot\widetilde{\boldsymbol{u}}^n) + \nabla\times(\nabla\times\widetilde{\boldsymbol{u}}^n)$. Using this identity in (74.6) yields

$$D_\tau^{(q)}\boldsymbol{u}^n + \mu\nabla\times(\nabla\times\widetilde{\boldsymbol{u}}^n) + \nabla\left(\beta_q\phi^n + p^{\star,n} - \mu\nabla\cdot\widetilde{\boldsymbol{u}}^n\right) = \boldsymbol{f}^n. \tag{74.11}$$

It is again possible to read this equation as a consistent approximation of the momentum balance equation if the quantity $\beta_q\phi^n + p^{\star,n} - \mu\nabla\cdot\widetilde{\boldsymbol{u}}^n$ is interpreted as an approximation of the pressure. An alternative way of writing the third substep of the projection algorithm thus consists of updating the pressure as

$$p^n := \beta_q\phi^n + p^{\star,n} - \mu\nabla\cdot\widetilde{\boldsymbol{u}}^n. \tag{74.12}$$

Henceforth, we consider the scheme composed of the three substeps (74.2)-(74.4)-(74.12). To understand why the modified scheme performs better than (74.2)-(74.4)-(74.5), we observe from (74.4) that $\nabla\times(\nabla\times\widetilde{\boldsymbol{u}}^n) = \nabla\times(\nabla\times\boldsymbol{u}^n)$. Therefore, (74.11) can be rewritten as

$$D_\tau^{(q)}\boldsymbol{u}^n + \mu\nabla\times(\nabla\times\boldsymbol{u}^n) + \nabla p^n = \boldsymbol{f}^n, \quad \nabla\cdot\boldsymbol{u}^n = 0, \quad \boldsymbol{u}^n_{|\partial D}\cdot\boldsymbol{n} = 0, \tag{74.13}$$

from which we deduce that $\boldsymbol{n}\cdot\nabla p^n_{|\partial D} = \boldsymbol{n}\cdot(\boldsymbol{f}^n - \mu\nabla\times(\nabla\times\boldsymbol{u}^n))_{|\partial D}$. Unlike (74.10), the pressure now satisfies a consistent pressure boundary condition. Hence, the splitting error is only due to the inexact tangential boundary condition on the velocity $\boldsymbol{u}^n$.

74.3.2 Stability and convergence properties

In order to give the reader some intuition on why the algorithm (74.2)-(74.4)-(74.12) is formally second-order accurate, we now consider a singular perturbation of the time-dependent Stokes problem that behaves like (74.2)-(74.4)-(74.12). Let us take $\beta_q := 1$ and $p^{\star,n} := p^{n-1}$ in (74.12), then setting $\epsilon := \tau$ and replacing $p^n - p^{\star,n}$ by $\epsilon\partial_t p$, the continuous version of (74.12) is $\epsilon\partial_t p = \phi - \mu\nabla\cdot\boldsymbol{u}$. Similarly, the continuous version of (74.4) is $\epsilon\Delta\phi = \nabla\cdot\boldsymbol{u}$, and the continuous version of (74.2) is $\partial_t\boldsymbol{u} - \nabla\cdot\mathbb{s}(\boldsymbol{u}) + \nabla p = \boldsymbol{f}$. This leads us to the following problem:

$$\partial_t\boldsymbol{u}^\varepsilon - \nabla\cdot\mathbb{s}(\boldsymbol{u}^\varepsilon) + \nabla p^\varepsilon = \boldsymbol{f}, \qquad \boldsymbol{u}^\varepsilon_{|\partial D} = \boldsymbol{0}, \qquad \boldsymbol{u}^\varepsilon(0) = \boldsymbol{u}_0, \tag{74.14a}$$

$$\nabla\cdot\boldsymbol{u}^\varepsilon - \varepsilon\Delta\phi^\varepsilon = 0, \qquad \boldsymbol{n}\cdot\nabla\phi^\varepsilon_{|\partial D} = 0, \tag{74.14b}$$

$$\varepsilon\partial_t p^\varepsilon = \phi^\varepsilon - \mu\nabla\cdot\boldsymbol{u}^\varepsilon, \qquad p^\varepsilon(0) = p(0). \tag{74.14c}$$

It turns out that the following lemma exhibits the essential properties of this singularly perturbed system, and its proof is the main guideline for the proof of Theorem 74.12 which essentially says that (74.2)-(74.4)-(74.12) is stable and second-order accurate in time.

Lemma 74.11 (Stability under perturbation). *Assume that the pair $(\boldsymbol{u}, p)$ is smooth enough in time and space, and that the regularity pickup for the Stokes problem is $s = 1$. There is a constant $c(p,T)$ such that the following holds true for all $\varepsilon > 0$,*

$$\|\nabla\cdot\boldsymbol{u}^\varepsilon\|_{L^\infty(J;L^2(D))} \le c(p,T)\mu^{-\frac{1}{2}}\varepsilon^{\frac{3}{2}}, \tag{74.15a}$$

$$\|\boldsymbol{u} - \boldsymbol{u}^\varepsilon\|_{L^2(J;\boldsymbol{L}^2(D))} \le c(p,T)\varepsilon^2. \tag{74.15b}$$

Proof. See [163, Lem. 3.1&3.2] and Exercise 74.4. □

Theorem 74.12 (Convergence: BDF2, $p^{\star,n} := p^{n-1}$). *Assume that the solution $(\boldsymbol{u}, p)$ to (73.2) is sufficiently smooth. Set $\boldsymbol{u}^0 := \boldsymbol{u}(0) = \boldsymbol{u}_0$ and $p^0 := p(0)$. Then the sequences $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ generated by (74.2)-(74.4)-(74.12), with* BDF2 *time stepping and the first-order pressure extrapolation $p^{\star,n} := p^{n-1}$, satisfy*

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \boldsymbol{u}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)} + \|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \widetilde{\boldsymbol{u}}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)} \le c(\boldsymbol{u},p,T)\tau^2,$$

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \widetilde{\boldsymbol{u}}_\tau\|_{\ell^2(J;\boldsymbol{V})} + \|\pi_\tau(p) - p_\tau\|_{\ell^2(J;Q)} \le c(\boldsymbol{u},p,T)\tau^{\frac{3}{2}}.$$

Proof. See Guermond and Shen [162, 163]. □

Remark 74.13 (Terminology and literature). In view of (74.13) where the operator $\nabla\times\nabla\times$ plays a key role, we refer to (74.2)-(74.4)-(74.12) as *pressure-correction scheme in rotational form*, and we refer to (74.2)-(74.4)-(74.5) as *pressure-correction scheme in standard form*. The method (74.2)-(74.4)-(74.12) has been first proposed in Timmermans et al. [275]. □

Remark 74.14 (Elimination of the projected velocity). As already mentioned in Remark 74.3, it is not necessary to compute the sequence of projected velocities $(\boldsymbol{u}^n)_{n\in\mathcal{N}_\tau}$ when implementing the above projection algorithms since these quantities can be algebraically eliminated. □

74.4 Finite element approximation

We now describe how the space semi-discrete setting from §73.2.1 can be used in conjunction with the pressure-projection algorithms introduced above. For the sake of brevity, we restrict ourselves

to the algorithm (74.7)-(74.8) where the velocity $(\boldsymbol{u}^n)_{n\in\mathcal{N}_\tau}$ has been eliminated, and we consider the first-order pressure extrapolation $p^{\star,n} := p^{n-1}$. Extensions to other variants of the method are straightforward. The discrete velocity spaces are $\boldsymbol{V}_h \subset \boldsymbol{V}$ and the discrete pressure spaces are $Q_h \subset L^2_*(D)$, for all $h \in \mathcal{H}$. We assume that Q_h is H^1-conforming. Although this hypothesis is not required by the approximation theory of the Stokes problem, it somewhat simplifies the presentation and the implementation of the method.

To avoid minor technical details, we initialize the algorithm by setting $\widetilde{\boldsymbol{u}}_h^0 := \boldsymbol{\mathcal{S}}_h^{\mathrm{v}}(\boldsymbol{u}_0, p(0))$ and $p_h^0 := \mathcal{S}_h^{\mathrm{p}}(\boldsymbol{u}_0, p(0))$, where the Stokes elliptic projections $(\boldsymbol{\mathcal{S}}_h^{\mathrm{v}}, \mathcal{S}_h^{\mathrm{p}})$ are defined in (73.9). We also set $\phi^0 := 0$. Then using $q := 1$ if $n = 1$ and $q := 2$ if $n \geq 2$, we consider the following sequence of problems:

(1) Find $\tilde{\boldsymbol{u}}_h^n \in \boldsymbol{V}_h$ such that for all $\boldsymbol{w}_h \in \boldsymbol{V}_h$,

$$(D_\tau^{(q)}\widetilde{\boldsymbol{u}}_h^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2} + a(\widetilde{\boldsymbol{u}}_h^n, \boldsymbol{w}_h) + b\Big(\boldsymbol{w}_h, p_h^{n-1} + \sum_{j\in\{1:q\}} \beta_{j-1}\phi^{n-j}\Big) = (\boldsymbol{f}^n, \boldsymbol{w}_h)_{\boldsymbol{L}^2}. \tag{74.16}$$

(2) Find $\phi_h^n \in Q_h$ such that

$$\tau(\nabla\phi_h^n, \nabla q_h)_{\boldsymbol{L}^2} = -(\nabla\cdot\widetilde{\boldsymbol{u}}_h^n, q_h)_{L^2}, \qquad \forall q_h \in Q_h. \tag{74.17}$$

(3) If the standard form of the algorithm is used, set

$$p_h^n := \beta_q\phi_h^n + p_h^{\star,n}, \tag{74.18}$$

whereas if the rotational form of the algorithm is used, set

$$p_h^n := \beta_q\phi_h^n + p_h^{\star,n} + \delta_h^n, \tag{74.19}$$

where $\delta_h^n \in Q_h$ is s.t. $(\delta_h^n, q_h)_{L^2} = (-\mu\nabla\cdot\widetilde{\boldsymbol{u}}_h^n, q_h)_{L^2}$ for all $q_h \in Q_h$.

Theorem 74.15 (Convergence). *Assume that the solution $(\boldsymbol{u}, p)$ to (73.2) is sufficiently smooth. Assume that full regularity pickup holds true. Then there is $c(\boldsymbol{u}, p, T)$ such that the sequences $\widetilde{\boldsymbol{u}}_\tau$, $\boldsymbol{u}_\tau$, p_τ generated by (74.16)-(74.17)-(74.18), with* BDF2 *time stepping and the first-order pressure extrapolation $p_h^{\star,n} := p_h^{n-1}$, satisfy*

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \boldsymbol{u}_{h\tau}\|_{\ell^\infty(J;\boldsymbol{L}^2)} + \|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \widetilde{\boldsymbol{u}}_{h\tau}\|_{\ell^\infty(J;\boldsymbol{L}^2)} \leq c(\boldsymbol{u}, p, T)(h^{k+1} + \tau^2),$$
$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u}) - \widetilde{\boldsymbol{u}}_{h\tau}\|_{\ell^2(J;\boldsymbol{V})} + \|\pi_\tau(p) - p_{h\tau}\|_{\ell^2(J;Q)} \leq c(\boldsymbol{u}, p, T)(h^k + \tau).$$

Proof. See Guermond [145, 143], Guermond and Quartapelle [160]. □

Although at the time of this writing no error analysis for the fully discrete scheme (74.16)-(74.17)-(74.19) has yet been published, it is generally believed, and confirmed by numerical tests, that with this scheme the second error estimate in Theorem 74.15 should be replaced by $c(h^k+\tau^{\frac{3}{2}})$.

Remark 74.16 (Discrete space for $\boldsymbol{u}_h^n$). Notice that the discrete velocity $\boldsymbol{u}_h^n$ has been eliminated from the algorithm (74.16)–(74.19). Recalling that we assumed that Q_h is H^1-conforming, the discrete space that is implicitly used in the projection step (74.17) for $\boldsymbol{u}_h^n$ is $\boldsymbol{V}_h + \nabla Q_h$. Hence, if needed, one recovers $\boldsymbol{u}_h^n$ by setting $\boldsymbol{u}_h^n = \widetilde{\boldsymbol{u}}_h^n - \tau\nabla\phi_h^n$. Another possibility when working with discontinuous pressures (i.e., Q_h is not H^1-conforming) is to replace the discrete Poisson problem (74.17) by a discrete version of the Darcy problem (74.4), where the discrete velocity $\boldsymbol{u}_h^n$ is sought in an $\boldsymbol{\mathcal{H}}$-conforming finite element space, e.g., built using Raviart–Thomas elements (see Chapter 14), and ϕ_h^n in a discontinuous finite element space. The reader is referred to Guermond [143] for further insight into these questions. □

Remark 74.17 (Inf-sup condition). Notice that the two discrete problems (74.16) and (74.17) can be solved in sequence and that none of them requires the inf-sup condition (73.4) (since they both involve a coercive bilinear form). One may be tempted to conclude that the scheme (74.16)-(74.17)-(74.18) (or (74.16)-(74.17)-(74.19)) is a way of solving the (Navier–)Stokes equations with finite elements without bothering about the inf-sup condition. This intuitive argument is false since the inf-sup condition must be satisfied for the above algorithms to yield the expected accuracy (see [143, 160] for the convergence proof). We also refer the reader to Burman et al. [76] for an analysis of projection methods using equal-order velocity and pressure finite element spaces together with fluctuation-based stabilization. □

Exercises

Exercise 74.1 (Remark 74.1). Prove the stability estimate in Remark 74.6. (*Hint*: adapt the proof of Lemma 74.5.)

Exercise 74.2 (Curl-div-grad identity). Let $d \in \{2,3\}$. Show that $\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}^2+\|\nabla\cdot\boldsymbol{v}\|_{L^2(D)}^2 = \|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}^2$ for all $\boldsymbol{v}\in\boldsymbol{H}_0^1(D)$. (*Hint:* use $-\Delta\boldsymbol{v} = -\nabla(\nabla\cdot\boldsymbol{v}) + \nabla\times(\nabla\times\boldsymbol{v})$.)

Exercise 74.3 (Inverse of the Stokes operator). Let $\boldsymbol{V} := \boldsymbol{H}_0^1(D)$, $\boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$, and $Q := L_*^2(D)$. The inverse of the Stokes operator $\boldsymbol{S} : \boldsymbol{H}^{-1}(D) \longrightarrow \boldsymbol{\mathcal{V}} := \{\boldsymbol{v}\in\boldsymbol{H}_0^1(D) \mid \nabla\cdot\boldsymbol{v} = 0\}$ is s.t. for all $\boldsymbol{f}\in\boldsymbol{V}'$, $\boldsymbol{S}(\boldsymbol{f})$ is the unique member of $\boldsymbol{\mathcal{V}}$ s.t. the following holds true for all $(\boldsymbol{w},q)\in\boldsymbol{V}\times Q$:

$$\begin{cases} 2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{f})),\mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)} - (r,\nabla\cdot\boldsymbol{w})_{L^2(D)} = \langle\boldsymbol{f},\boldsymbol{w}\rangle_{\boldsymbol{V}',\boldsymbol{V}},\\ (q,\nabla\cdot\boldsymbol{S}(\boldsymbol{f}))_{L^2(D)} = 0,\end{cases}$$

where $\langle\cdot,\cdot\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ denotes the duality pairing between $\boldsymbol{V}'$ and $\boldsymbol{V}$. Recall that $\mu\|\boldsymbol{S}(\boldsymbol{f})\|_{\boldsymbol{V}} + \|r\|_{L^2} \le c\|\boldsymbol{f}\|_{\boldsymbol{H}^{-1}}$ for all $\boldsymbol{f}\in\boldsymbol{H}^{-1}(D)$ with $\|\boldsymbol{w}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{w})\|_{\mathbb{L}^2(D)}$. We assume that D is such that the following regularity property holds true: $\mu|\boldsymbol{S}(\boldsymbol{f})|_{\boldsymbol{H}^2} + |r|_{H^1} \le c\|\boldsymbol{f}\|_{\boldsymbol{L}^2}$ for all $\boldsymbol{f}\in\boldsymbol{L}^2(D)$. (i) Show that $2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} = \|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2$ for all $\boldsymbol{v}\in\boldsymbol{\mathcal{V}}$. (*Hint*: recall that the duality pairing $\langle\cdot,\cdot\rangle_{\boldsymbol{V}',\boldsymbol{V}}$ is an extension of the $\boldsymbol{L}^2$-inner product.) (ii) Show that for all $\gamma\in(0,1)$, there is $c(\gamma)$ such that for all $\boldsymbol{v}$ in $\boldsymbol{V}$, $2\mu(\mathbb{e}(\boldsymbol{S}(\boldsymbol{v})),\mathbb{e}(\boldsymbol{v}))_{\mathbb{L}^2} \ge (1-\gamma)\|\boldsymbol{v}\|_{\boldsymbol{L}^2}^2 - c(\gamma)\|\boldsymbol{v}-\boldsymbol{v}^\star\|_{\boldsymbol{L}^2}^2$ for all $\boldsymbol{v}^\star\in\boldsymbol{\mathcal{H}}$. (*Hint*: integrate by parts the pressure term.) (iii) Show that the map $\boldsymbol{V}'\ni\boldsymbol{v}\mapsto|\boldsymbol{v}|_\star := \langle\boldsymbol{v},\boldsymbol{S}(\boldsymbol{v})\rangle_{\boldsymbol{V}',\boldsymbol{V}}^{\frac12}$ defines a seminorm on $\boldsymbol{V}'$. Prove that $|\boldsymbol{v}|_\star \le (2\mu)^{-\frac12}\|\boldsymbol{v}\|_{\boldsymbol{V}'}$ for all $\boldsymbol{v}\in\boldsymbol{V}'$. *Note*: there does not exist any constant c so that $(2\mu)^{-\frac12}\|\boldsymbol{v}\|_{\boldsymbol{V}'}\le c|\boldsymbol{v}|_\star$ for all $\boldsymbol{v}\in\boldsymbol{H}^{-1}(D)$, i.e., $|\cdot|_\star$ is not a norm on $\boldsymbol{H}^{-1}(D)$; see Guermond [142, Thm. 4.1] and Guermond and Salgado [161, Thm. 32]. The inverse of the Stokes operator is used in Exercise 74.4 to prove Lemma 74.11.

Exercise 74.4 (Lemma 74.11). Consider the perturbed system (74.14), and set $\boldsymbol{e} := \boldsymbol{u}^\varepsilon - \boldsymbol{u}$ and $q := p^\varepsilon - p$. (i) Write the PDE system solved by the pair $(\boldsymbol{e},q)$ and show that

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\partial_t\boldsymbol{e}\|_{\boldsymbol{L}^2}^2 + 2\mu\|\partial_t\boldsymbol{e}\|_{\boldsymbol{V}}^2 + \frac12\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2 + \frac12\varepsilon\mu\frac{\mathrm{d}}{\mathrm{d}t}\|\Delta\phi^\varepsilon\|_{L^2}^2 = \varepsilon\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\partial_t p,\nabla\phi^\varepsilon)_{\boldsymbol{L}^2} - \varepsilon(\nabla\partial_{tt}p,\nabla\phi^\varepsilon)_{\boldsymbol{L}^2},$$

where we recall that $\boldsymbol{V} := \boldsymbol{H}_0^1(D)$ and $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2}$. (ii) Prove that $\|\nabla\phi^\varepsilon(t)\|_{\boldsymbol{L}^2}^2 \le c(p,T)\varepsilon^2$ for all $t\in J$. (*Hint*: use Gronwall's lemma from Exercise 65.3.) Conclude that $\|\nabla\cdot\boldsymbol{u}^\varepsilon\|_{L^\infty(J;L^2(D))}^2 \le c(p,T)\mu^{-1}\varepsilon^3$. (iii) Show that $\|\boldsymbol{e} - \boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{e})\|_{\boldsymbol{L}^2}^2 = \varepsilon^2\|\nabla\phi^\varepsilon\|_{\boldsymbol{L}^2}^2$, where the

Leray projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}$ is defined in Lemma 74.1. Deduce from the above estimates that $\|\boldsymbol{u} - \boldsymbol{u}^\varepsilon\|_{L^2(J;\boldsymbol{L}^2(D))} \le c(p,T)\varepsilon^2$. (*Hint*: use the lower bound from Step (ii) of Exercise 74.3.)

Exercise 74.5 (Gauge-Uzawa). (i) Write the pressure-correction algorithm in rotational form using BDF1, $p^{\star,n} := p^{n-1}$, and the sequences $\widetilde{\boldsymbol{u}}_\tau \in (\boldsymbol{V})^N$, $\boldsymbol{u}_\tau \in (\boldsymbol{\mathcal{H}})^N$, $\phi_\tau \in (Q)^N$, $p_\tau \in (Q)^N$. (ii) Consider the sequences $\widetilde{\boldsymbol{v}}_\tau \in (\boldsymbol{V})^N$, $\boldsymbol{v}_\tau \in (\boldsymbol{V})^N$, $r_\tau \in (Q)^N$, $q_\tau \in (Q)^N$, $\psi_\tau \in (Q)^N$, generated by the following algorithm (called gauge-Uzawa in the literature, see Nochetto and Pyo [230]): Set $\boldsymbol{v}^0 := \boldsymbol{v}_0$, $r^0 := 0$, $q^0 = \psi^0 := p(0)$, then solve for all $n \in \mathcal{N}_\tau$,

$$
\begin{aligned}
&\frac{\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}}{\tau} - \mu\Delta\widetilde{\boldsymbol{v}}^n + \nabla q^{n-1} = \boldsymbol{f}^n, \quad \widetilde{\boldsymbol{v}}^n_{|\partial D} = \boldsymbol{0},\\
&\boldsymbol{v}^n + \tau\nabla\psi^n = \widetilde{\boldsymbol{v}}^n + \tau\nabla\psi^{n-1}, \quad \nabla\cdot\boldsymbol{v}^n = 0, \quad \boldsymbol{v}^n_{|\partial D}\cdot\boldsymbol{n} = 0,\\
&r^n = r^{n-1} - \nabla\cdot\widetilde{\boldsymbol{v}}^n, \qquad\qquad q^n = \psi^n + \mu r^n.
\end{aligned}
$$

Recalling that $(\delta_\tau\psi_\tau)^n := \frac{\psi^n - \psi^{n-1}}{\tau}$ for all $n \in \mathcal{N}_\tau$, show that the sequences $(\widetilde{\boldsymbol{v}}_\tau, \boldsymbol{v}_\tau, \tau\delta_\tau\psi_\tau, q_\tau)$ and $(\widetilde{\boldsymbol{u}}_\tau, \boldsymbol{u}_\tau, \phi_\tau, p_\tau)$ are equal (i.e., the gauge-Uzawa and the pressure-correction method in rotational form are identical). (*Hint*: write $q^n = q^{n-1} + \psi^n - \psi^{n-1} + \mu(r^n - r^{n-1})$.) (iii) Show that for all $n \in \mathcal{N}_\tau$,

$$
\begin{aligned}
\|\boldsymbol{v}^n\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla\psi^n\|^2_{\boldsymbol{L}^2} + \mu\tau\|r^n\|^2_{L^2} + \|\widetilde{\boldsymbol{v}}^n - \boldsymbol{v}^{n-1}\|^2_{\boldsymbol{L}^2} + \frac{1}{2}\mu\tau\|\nabla\widetilde{\boldsymbol{v}}^n\|^2_{\mathbb{L}^2}&\\
\le \|\boldsymbol{v}^{n-1}\|^2_{\boldsymbol{L}^2} + \tau^2\|\nabla\psi^{n-1}\|^2_{\boldsymbol{L}^2} + \mu\tau\|r^{n-1}\|^2_{L^2} + \rho\tau\|\boldsymbol{f}^n\|^2_{\boldsymbol{L}^2},&
\end{aligned}
$$

with the time scale $\rho := \frac{2}{C_{\rm PS}^2}\frac{\ell_D^2}{\mu}$. (*Hint*: test the momentum equation with $2\tau\widetilde{\boldsymbol{v}}^n$, square the second equation, square the third equation and scale the result by $\mu\tau$, and add the results.)

Chapter 75

Artificial compressibility

In this chapter, we study a time-stepping technique for the time-dependent Stokes equations based on an artificial compressibility perturbation of the mass conservation equation. This technique presents some advantages with respect to the projection methods studied in Chapter 74. It avoids solving a Poisson equation at each time step, and it can be extended to high order in a rather straightforward manner. To obtain $\mathcal{O}(\tau^k)$ accuracy, $k \geq 1$, the cost per time step is that of solving k vector-valued parabolic equations implicitly (notice that solving a Poisson equation is more expensive than taking an implicit time step for a parabolic equation). The material of this chapter is adapted from [150, 151, 152].

75.1 Stability under compressibility perturbation

As in the previous chapter, we assume for simplicity that homogeneous Dirichlet boundary conditions are enforced on the velocity over the entire boundary and that the viscosity μ is constant. Recall that $\boldsymbol{V} := \boldsymbol{H}^1_0(D)$ and $\|\boldsymbol{v}\|_{\boldsymbol{V}} := \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$. The main idea of the *artificial compressibility method* is to replace the time-dependent Stokes equations

$$\begin{cases} \partial_t \boldsymbol{u} - \nabla\cdot\mathbb{s}(\boldsymbol{u}) + \nabla p = \boldsymbol{f}, \quad \boldsymbol{u}(0) = \boldsymbol{u}_0, \quad \boldsymbol{u}_{|\partial D} = \boldsymbol{0}, \\ \nabla\cdot\boldsymbol{u} = 0, \end{cases} \tag{75.1}$$

by the following perturbed problem:

$$\begin{cases} \partial_t \boldsymbol{u}^\epsilon - \nabla\cdot\mathbb{s}(\boldsymbol{u}^\epsilon) + \nabla p^\epsilon = \boldsymbol{f}, \quad \boldsymbol{u}^\epsilon(0) = \boldsymbol{u}_0, \quad \boldsymbol{u}^\epsilon_{|\partial D} = \boldsymbol{0}, \\ \epsilon\partial_t p^\epsilon + \nabla\cdot\boldsymbol{u}^\epsilon = 0, \qquad\qquad\qquad p^\epsilon(0) = p(0), \end{cases} \tag{75.2}$$

with the perturbation parameter $\epsilon := \overline{\epsilon}\frac{\rho}{\mu}$, where $\rho := C_{\mathrm{KPS}}^{-2}\frac{\ell_D^2}{\mu}$ is the time scale and $\overline{\epsilon} > 0$ is a positive (nondimensional) number s.t. $\overline{\epsilon} \ll 1$. In the context of the time discretization, we shall set $\epsilon := \frac{\tau}{\lambda}$ with $\lambda := \lambda_0\mu$ and λ_0 is a positive (nondimensional) number of order 1 (so that $\overline{\epsilon} = \frac{1}{\lambda_0}\frac{\tau}{\rho}$). Notice that in (75.2) the compressibility perturbation involves the time derivative of the perturbed pressure, so that an initial condition for p^ϵ is needed. In what follows, we assume that the solution to (75.1) is smooth enough so that the initial pressure $p(0)$ is available (see Remark 74.4).

The cornerstone of the analysis of the artificial compressibility method is the stability of the time-dependent Stokes equations under a compressibility perturbation. Consider the following

abstract problem: Let $\boldsymbol{k} \in L^2(J;\boldsymbol{V}')$ and $g \in H^1(J;L^2_*(D))$, and let $(\boldsymbol{v},q)$ solve

$$\begin{cases} \partial_t \boldsymbol{v} - \nabla\cdot\mathbb{s}(\boldsymbol{v}) + \nabla q = \boldsymbol{k}, & \boldsymbol{v}(0) = \boldsymbol{v}_0, \quad \boldsymbol{v}_{|\partial D} = \boldsymbol{0}, \\ \epsilon\partial_t q + \nabla\cdot\boldsymbol{v} = g, & q(0) = q_0, \end{cases} \tag{75.3}$$

with the initial data $\boldsymbol{v}_0 \in \boldsymbol{L}^2(D)$ and $q_0 \in L^2(D)$.

Lemma 75.1 (Stability under perturbation). *Let $(\boldsymbol{v},q)$ solve (75.3).* (i) *There is c s.t. the following a priori estimate holds true for all $\epsilon > 0$:*

$$\frac{1}{2}\|\boldsymbol{v}\|^2_{L^\infty(J;\boldsymbol{L}^2)} + \epsilon\|q\|^2_{L^\infty(J;\boldsymbol{L}^2)} + \mu\|\boldsymbol{v}\|^2_{L^2(J;\boldsymbol{V})} \le 4\|\boldsymbol{v}_0\|^2_{\boldsymbol{L}^2} + 2\epsilon\|q_0\|^2_{L^2} + c\left(\mu^{-1}\|\boldsymbol{k}\|^2_{L^2(J;\boldsymbol{V}')} + \mu\|g\|^2_{H^1(J;L^2)}\right). \tag{75.4}$$

(ii) *If in addition $\boldsymbol{k} \in H^1(J;\boldsymbol{L}^2(D))$, $g \in H^2(J;L^2_*(D))$, and $(\boldsymbol{v}_0,q_0)$ is smooth enough so that the momentum and the mass equations in (75.3) hold true at the initial time, i.e., $\partial_t\boldsymbol{v}(0) = \boldsymbol{k}(0) + \nabla\cdot\mathbb{s}(\boldsymbol{v}_0) - \nabla q_0 \in \boldsymbol{L}^2(D)$ and $\epsilon\partial_t q(0) = g(0) - \nabla\cdot\boldsymbol{v}_0 \in L^2(D)$, then we have*

$$\|q\|^2_{L^2(J;L^2)} \le c\left(\mu\Big(\rho^2\big(\|\partial_t\boldsymbol{v}(0)\|^2_{\boldsymbol{L}^2} + \epsilon\|\partial_t q(0)\|^2_{L^2}\big) + \|\boldsymbol{v}_0\|^2_{\boldsymbol{L}^2} + \epsilon\|q_0\|^2_{L^2}\Big) + \|\boldsymbol{k}\|^2_{H^1(J;\boldsymbol{V}')} + \mu^2\|g\|^2_{H^2(J;L^2)}\right). \tag{75.5}$$

Proof. See Exercise 75.1. □

75.2 First-order artificial compressibility

In this section, we construct an artificial compressibility method that is first-order accurate in time. This is done by using the implicit Euler time-stepping scheme to discretize the momentum and the mass conservation equations in (75.2). Let τ be the time step. Let $\lambda := \lambda_0\mu$, where λ_0 is a positive (nondimensional) number of order 1. Let us set $\epsilon := \frac{\tau}{\lambda}$. To initialize the first-order artificial compressibility method, we set $\boldsymbol{u}^0 := \boldsymbol{u}_0$ and $p^0 := p(0)$. Then for all $n \in \mathcal{N}_\tau$, the pair $(\boldsymbol{u}^n, p^n)$ is computed by using the implicit Euler time-stepping scheme in (75.2):

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{u}^n - \boldsymbol{u}^{n-1}) - \nabla\cdot\mathbb{s}(\boldsymbol{u}^n) + \nabla p^n = \boldsymbol{f}^n, \quad \boldsymbol{u}^n_{|\partial D} = \boldsymbol{0}, \\ \dfrac{1}{\lambda}(p^n - p^{n-1}) + \nabla\cdot\boldsymbol{u}^n = 0, \end{cases} \tag{75.6}$$

where $(\boldsymbol{u}^{n-1}, p^{n-1})$ is known from the previous step or the initial condition. Notice that we replaced $\frac{\epsilon}{\tau}$ by $\frac{1}{\lambda}$. A crucial observation is that the velocity and the pressure are uncoupled: the second equation gives $p^n = p^{n-1} - \lambda\nabla\cdot\boldsymbol{u}^n$, and substituting the value of p^n in the first equation we obtain

$$\begin{cases} \boldsymbol{u}^n - \tau\big(\nabla\cdot\mathbb{s}(\boldsymbol{u}^n) + \lambda\nabla\nabla\cdot\boldsymbol{u}^n\big) = \boldsymbol{u}^{n-1} + \tau(\boldsymbol{f}^n - \nabla p^{n-1}), \\ p^n = p^{n-1} - \lambda\nabla\cdot\boldsymbol{u}^n. \end{cases} \tag{75.7}$$

The main advantage of the above technique with respect to the saddle point problem (73.17) is that the complexity of solving (75.6) is the same as solving one implicit step of a parabolic problem on the velocity.

To analyze the stability and convergence properties of the first-order artificial compressibility method, we set $(\boldsymbol{u}_\tau, p_\tau) := (\boldsymbol{u}^n, p^n)_{n\in\mathcal{N}_\tau}$ and recall the notation $\boldsymbol{\pi}_\tau(\boldsymbol{u}) := (\boldsymbol{u}(t_n))_{n\in\mathcal{N}_\tau}$, where $(\boldsymbol{u}, p)$ solves (75.1). We also use the notation $\delta_\tau \boldsymbol{v}_\tau := (\delta_\tau \boldsymbol{v}^n)_{n\in\mathcal{N}_\tau}$ with $\delta_\tau \boldsymbol{v}^n := \frac{\boldsymbol{v}^n - \boldsymbol{v}^{n-1}}{\tau}$ for all $n \in \mathcal{N}_\tau$. Recall the norms $\|\phi_\tau\|^2_{\ell^2(J;B)} := \sum_{n\in\mathcal{N}_\tau} \tau\|\phi^n\|_B^2$ and $\|\phi_\tau\|_{\ell^\infty(J;V)} := \max_{n\in\mathcal{N}_\tau} \|\phi^n\|_B$ (notice that the maximum is taken over $n \in \mathcal{N}_\tau$ in the norm $\|\phi_\tau\|_{\ell^2(J;B)}$, whereas it is taken over $n \in \overline{\mathcal{N}}_\tau := \{0\!:\!N\}$ in the norm $\|\phi_\tau\|_{\ell^\infty(\overline{J};B)}$).

Lemma 75.2 (Stability). *Let $\boldsymbol{k} \in C^1(\overline{J}; \boldsymbol{L}^2(D))$, $g \in C^2(\overline{J}; L^2_*(D))$, $\boldsymbol{u}^0 \in \boldsymbol{L}^2(D)$, and $p^0 \in L^2(D)$. Assume that $\tau \le \frac14\rho$. With the notation $\boldsymbol{k}^n := \boldsymbol{k}(t_n)$ and $g^n := g(t_n)$ for all $n \in \mathcal{N}_\tau$, let $(\boldsymbol{u}_\tau, p_\tau)$ be the time sequence s.t.*

$$\left\{\begin{aligned} &\frac{1}{\tau}(\boldsymbol{u}^n - \boldsymbol{u}^{n-1}) - \nabla\cdot\mathbb{s}(\boldsymbol{u}^n) + \nabla p^n = \boldsymbol{k}^n, \quad \boldsymbol{u}^n_{|\partial D} = \boldsymbol{0},\\ &\frac{1}{\lambda}(p^n - p^{n-1}) + \nabla\cdot\boldsymbol{u}^n = g^n. \end{aligned}\right. \tag{75.8}$$

(i) *Letting $J_* := (t_1, T)$, there is c such that for all $\tau > 0$,*

$$\begin{aligned}\|\boldsymbol{u}_\tau\|^2_{\ell^\infty(J;\boldsymbol{L}^2)} + \mu\|\boldsymbol{u}_\tau\|^2_{\ell^2(J;\boldsymbol{V})} \le c\, e^{\frac{4T}{\rho}}\Big(&\|\boldsymbol{u}^0\|^2_{\boldsymbol{L}^2} + \frac{\tau}{\mu}\|p^0\|^2_{L^2} + \rho\|\boldsymbol{k}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)}\\ &+ \mu(T+\rho)\|g_\tau\|^2_{\ell^\infty(J;L^2)} + \mu\rho^2\|\partial_t g\|^2_{L^2(J_*;L^2)}\Big).\end{aligned} \tag{75.9}$$

(ii) *Letting $J_{**} := (t_2, T)$ and $\|\phi_\tau\|_{\ell^\infty(J_*;B)} := \max_{n\in\{2:N\}}\|\phi^n\|_B$, we have*

$$\begin{aligned}\|p_\tau\|^2_{\ell^2(J;L^2)} \le c\,\mu e^{\frac{4T}{\rho}}\Big(&\|\boldsymbol{u}^0\|^2_{\boldsymbol{L}^2} + \rho^2\|\delta_\tau\boldsymbol{u}^1\|^2_{\boldsymbol{L}^2} + \frac{\tau}{\mu}\big(\|p^0\|^2_{L^2} + \rho^2\|\delta_\tau p^1\|^2_{L^2}\big)\\ &+ \rho\big(\|\boldsymbol{k}_\tau\|^2_{\ell^2(J;\boldsymbol{L}^2)} + \rho^2\|\partial_t\boldsymbol{k}\|^2_{L^2(J_*;\boldsymbol{L}^2)}\big)\\ &+ \mu(T+\rho)\big(\|g_\tau\|^2_{\ell^\infty(J;L^2)} + \rho^2\|\delta_\tau g_\tau\|^2_{\ell^\infty(J_*;L^2)}\big)\\ &+ \mu\rho^2\big(\|\partial_t g\|^2_{L^2(J_*;L^2)} + \rho^2\|\partial_{tt} g\|^2_{L^2(J_*;L^2)}\big)\Big).\end{aligned} \tag{75.10}$$

Proof. We only prove the estimate (75.9) and refer the reader to Exercise 75.2 for the proof of (75.10); see also Shen [255, Prop. 5.1]. Testing the momentum equation in (75.8) with $2\tau\boldsymbol{u}^n$ and the mass equation with $2\tau p^n$, adding the two results, using the coercivity of the bilinear form $a(\boldsymbol{v}, \boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2}$ and Young's inequality to estimate $(\boldsymbol{k}^n, \boldsymbol{u}^n)_{\boldsymbol{L}^2}$, and dropping the nonnegative terms $\|\boldsymbol{u}^n - \boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2}$ and $\frac{\tau}{\lambda}\|p^n - p^{n-1}\|^2_{L^2}$ from the left-hand side leads to

$$\|\boldsymbol{u}^n\|^2_{\boldsymbol{L}^2} - \|\boldsymbol{u}^{n-1}\|^2_{\boldsymbol{L}^2} + \frac{\tau}{\lambda}\|p^n\|^2_{L^2} - \frac{\tau}{\lambda}\|p^{n-1}\|^2_{L^2} + 3\mu\tau\|\boldsymbol{u}^n\|^2_{\boldsymbol{V}} \le \rho\tau\|\boldsymbol{k}^n\|^2_{\boldsymbol{L}^2} + 2\tau(p^n, g^n)_{L^2}.$$

Owing to Lemma 53.9, there exists β_D and a linear map $\boldsymbol{w} : L^2_*(D) \to \boldsymbol{V}$ s.t. for all $g \in L^2_*(D)$, $\nabla\cdot(\boldsymbol{w}(g)) = g$ and $\beta_D\|\boldsymbol{w}(g)\|_{\boldsymbol{V}} \le \|g\|_{L^2(D)}$ (this map is a right inverse of the divergence operator; see Lemma C.44). Setting $\boldsymbol{w}^n := \boldsymbol{w}(g^n)$ and using Young's inequality and $\|\boldsymbol{w}^n\|_{\boldsymbol{L}^2} \le (\rho\mu)^{\frac12}\|\boldsymbol{w}^n\|_{\boldsymbol{V}}$, we infer that

$$\begin{aligned}(p^n, g^n)_{L^2} = (p^n, \nabla\cdot\boldsymbol{w}^n)_{L^2} &= (-\boldsymbol{k}^n + \tfrac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau} - \nabla\cdot\mathbb{s}(\boldsymbol{u}^n), \boldsymbol{w}^n)_{\boldsymbol{L}^2}\\ &\le \rho^{\frac12}\|\boldsymbol{k}^n\|_{\boldsymbol{L}^2}\mu^{\frac12}\|\boldsymbol{w}^n\|_{\boldsymbol{V}} + 2\mu\|\boldsymbol{u}^n\|_{\boldsymbol{V}}\|\boldsymbol{w}^n\|_{\boldsymbol{V}} + (\tfrac{\boldsymbol{u}^n - \boldsymbol{u}^{n-1}}{\tau}, \boldsymbol{w}^n)_{\boldsymbol{L}^2}\\ &\le \frac12\rho\|\boldsymbol{k}^n\|^2_{\boldsymbol{L}^2} + \mu\|\boldsymbol{u}^n\|^2_{\boldsymbol{V}} + \frac32\mu\|\boldsymbol{w}^n\|^2_{\boldsymbol{V}} + \frac{1}{\tau}(\boldsymbol{u}^n, \boldsymbol{w}^n)_{\boldsymbol{L}^2} - \frac{1}{\tau}(\boldsymbol{u}^{n-1}, \boldsymbol{w}^n)_{\boldsymbol{L}^2}.\end{aligned}$$

Inserting this bound into the previous estimate shows that

$$\begin{aligned}&\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2-\|\boldsymbol{u}^{n-1}\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^n\|_{L^2}^2-\frac{\tau}{\lambda}\|p^{n-1}\|_{L^2}^2+\mu\tau\|\boldsymbol{u}^n\|_{\boldsymbol{V}}^2\\&\le 2\rho\tau\|\boldsymbol{k}^n\|_{\boldsymbol{L}^2}^2+3\mu\tau\|\boldsymbol{w}^n\|_{\boldsymbol{V}}^2+2(\boldsymbol{u}^n,\boldsymbol{w}^n)_{\boldsymbol{L}^2}-2(\boldsymbol{u}^{n-1},\boldsymbol{w}^n)_{\boldsymbol{L}^2}.\end{aligned}$$

We write the above estimate for any index $l\in\{1\!:\!n\}$ and sum over l. Using the notation $\|\phi_\tau\|_{\ell^2((0,t_n);B)}^2:=\sum_{l\in\{1:n\}}\tau\|\phi^l\|_B^2$ for a time sequence $\phi_\tau:=(\phi^n)_{n\in\mathcal{N}_\tau}\in(B)^N$ and a Banach space B equipped with the norm $\|\cdot\|_B$ yields

$$\begin{aligned}&\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^n\|_{L^2}^2+\mu\|\boldsymbol{u}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\\&\le\|\boldsymbol{u}^0\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^0\|_{L^2}^2+2\rho\|\boldsymbol{k}_\tau\|_{\ell^2((0,t_n);\boldsymbol{L}^2)}^2+3\mu\|\boldsymbol{w}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\\&\quad+2(\boldsymbol{u}^n,\boldsymbol{w}^n)_{\boldsymbol{L}^2}-2(\boldsymbol{u}^0,\boldsymbol{w}^1)_{\boldsymbol{L}^2}-\sum_{l\in\{1:n-1\}}2(\boldsymbol{u}^l,\boldsymbol{w}^{l+1}-\boldsymbol{w}^l)_{\boldsymbol{L}^2},\end{aligned}$$

where we used that $(\boldsymbol{u}^l,\boldsymbol{w}^l)_{\boldsymbol{L}^2}-(\boldsymbol{u}^{l-1},\boldsymbol{w}^l)_{\boldsymbol{L}^2}=(\boldsymbol{u}^l,\boldsymbol{w}^l)_{\boldsymbol{L}^2}-(\boldsymbol{u}^{l-1},\boldsymbol{w}^{l-1})_{\boldsymbol{L}^2}-(\boldsymbol{u}^{l-1},\boldsymbol{w}^l-\boldsymbol{w}^{l-1})_{\boldsymbol{L}^2}$. Owing to the Cauchy–Schwarz inequality, we infer that

$$\begin{aligned}&\frac12\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^n\|_{L^2}^2+\mu\|\boldsymbol{u}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\\&\le2\|\boldsymbol{u}^0\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^0\|_{L^2}^2+2\rho\|\boldsymbol{k}_\tau\|_{\ell^2((0,t_n);\boldsymbol{L}^2)}^2+3\mu\|\boldsymbol{w}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\\&+2\|\boldsymbol{w}^n\|_{\boldsymbol{L}^2}^2+\|\boldsymbol{w}^1\|_{\boldsymbol{L}^2}^2+\sum_{l\in\{1:n-1\}}\frac{\rho}{\tau}\|\boldsymbol{w}^{l+1}-\boldsymbol{w}^l\|_{\boldsymbol{L}^2}^2+\sum_{l\in\{1:n-1\}}\frac{\tau}{\rho}\|\boldsymbol{u}^l\|_{\boldsymbol{L}^2}^2.\end{aligned}$$

We observe that $\|\boldsymbol{w}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\le\beta_D^{-2}\|g_\tau\|_{\ell^2((0,t_n);L^2)}^2\le\beta_D^{-2}T\|g_\tau\|_{\ell^\infty(J;L^2)}^2$. Moreover, using that $\|\boldsymbol{w}^l\|_{\boldsymbol{L}^2}^2\le\mu\rho\|\boldsymbol{w}^l\|_{\boldsymbol{V}}^2\le\beta_D^{-2}\mu\rho\|g^l\|_{L^2}^2$ for all $l\in\mathcal{N}_\tau$, we obtain $2\|\boldsymbol{w}^n\|_{\boldsymbol{L}^2}^2+\|\boldsymbol{w}^1\|_{\boldsymbol{L}^2}^2\le3\beta_D^{-2}\mu\rho\|g_\tau\|_{\ell^\infty(J;L^2)}^2$. Finally, observing that $\|\boldsymbol{w}^{l+1}-\boldsymbol{w}^l\|_{\boldsymbol{V}}=\|\boldsymbol{w}(g^{l+1}-g^l)\|_{\boldsymbol{V}}\le\beta_D^{-1}\|\int_{J_{l+1}}\partial_t g\mathrm{d}t\|_{L^2}$ owing to the linearity of the map $\boldsymbol{w}$, we infer that $\|\boldsymbol{w}^{l+1}-\boldsymbol{w}^l\|_{\boldsymbol{V}}\le\beta_D^{-1}\tau^{\frac12}\|\partial_t g\|_{L^2(J_{l+1};L^2)}$ owing to the Cauchy–Schwarz inequality. These bounds give

$$\begin{aligned}&\frac12\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^n\|_{L^2}^2+\mu\|\boldsymbol{u}_\tau\|_{\ell^2((0,t_n);\boldsymbol{V})}^2\\&\le2\|\boldsymbol{u}^0\|_{\boldsymbol{L}^2}^2+\frac{\tau}{\lambda}\|p^0\|_{L^2}^2+2\rho\|\boldsymbol{k}_\tau\|_{\ell^2((0,t_n);\boldsymbol{L}^2)}^2+3\beta_D^{-2}\mu(T+\rho)\|g_\tau\|_{\ell^\infty(J;L^2)}^2\\&\quad+\beta_D^{-2}\mu\rho^2\|\partial_t g\|_{L^2((t_1,t_n);L^2)}^2+\sum_{l\in\{1:n-1\}}\frac{\tau}{\rho}\|\boldsymbol{u}^l\|_{\boldsymbol{L}^2}^2.\end{aligned}$$

To conclude, we apply the discrete Gronwall lemma from Exercise 68.3 with $\gamma:=\frac{2\tau}{\rho}$, observe that $\frac{1}{1-\gamma}\le e^{2\gamma}$ since $\gamma\in(0,\frac12)$ by assumption, drop the nonnegative term $\frac{\tau}{\lambda}\|p^n\|_{L^2}^2$, and use that $\lambda:=\lambda_0\mu$. □

We now establish a convergence result for (75.7). The generic constant $c(\boldsymbol{u},p,T)$ depends on $\boldsymbol{u}$, p, and T, but is independent of τ.

Proposition 75.3 (Convergence). *Let $(\boldsymbol{u}_\tau,p_\tau)$ solve (75.7). Let $(\boldsymbol{u},p)$ solve (75.1). Assume that $\boldsymbol{u}\in H^2(J;\boldsymbol{L}^2(D))$ and $p\in H^2(J;L^2_*(D))$. Assume that $\tau\le\frac14\rho$.* (i) *There is $c(\boldsymbol{u},p,T)$ s.t. for all $\tau>0$,*

$$\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\boldsymbol{u}_\tau\|_{\ell^\infty(J;\boldsymbol{L}^2)}+\mu^{\frac12}\|\boldsymbol{\pi}_\tau(\boldsymbol{u})-\boldsymbol{u}_\tau\|_{\ell^2(J;\boldsymbol{V})}\le c(\boldsymbol{u},p,T)\tau.\tag{75.11}$$

(ii) *If* $\boldsymbol{u} \in H^3(J;\boldsymbol{L}^2(D))$ *and* $p \in H^3(J;L^2_*(D))$*, there is* $c'(\boldsymbol{u},p,T)$ *s.t.*

$$\|\pi_\tau(p) - p_\tau\|_{\ell^2(J,L^2)} \le c'(\boldsymbol{u},p,T)\tau^{\frac{1}{2}}. \tag{75.12}$$

Proof. See Shen [255, Prop. 5.1] and Exercise 75.3. (By proceeding as in Shen [255, Lem. 3.2], one can also prove that $\sum_{n\in\mathcal{N}_\tau} \min(\frac{t_n}{\rho},1)\|p(t_n)-p^n\|_{L^2}^2 \le c''(\boldsymbol{u},p,T)\tau$.) □

Remark 75.4 (Grad-div stabilization). The weak form of the first equation in (75.7) amounts to

$$(\boldsymbol{u}^n,\boldsymbol{w})_{\boldsymbol{L}^2} + \tau\big(a(\boldsymbol{u}^n,\boldsymbol{w}) + \lambda(\nabla\cdot\boldsymbol{u}^n,\nabla\cdot\boldsymbol{w})_{L^2}\big) = (\boldsymbol{g}^n,\boldsymbol{w})_{\boldsymbol{L}^2},$$

for all $\boldsymbol{w}\in\boldsymbol{V}$, with $\boldsymbol{g}^n := \boldsymbol{u}^{n-1} + \tau(\boldsymbol{f}^n - \nabla p^{n-1})$, i.e., artificial compressibility adds a grad-div stabilization to the momentum equation. □

Remark 75.5 (Literature). The artificial compressibility regularization can be traced back in the Russian literature to the group of Yanenko [281], [287, §8.2]. The variant (75.3) has also been proposed by Chorin [82, Eq. (3)] and analyzed by Temam [271, Eq. (0.3)]. □

75.3 Higher-order artificial compressibility

In this section, we use a Taylor series argument and a bootstrapping technique to construct a higher-order version in time of the artificial compressibility method introduced in §75.2. We start with the Taylor series argument. The method is general and can be deployed to any approximation order, but for simplicity we exemplify it for the third order. Let us set $\boldsymbol{u}_l := \partial_t^l\boldsymbol{u}$ for all $l\in\mathbb{N}$ (with the convention that $\boldsymbol{u}_0 := \boldsymbol{u}$), and let us set $\boldsymbol{u}_l^n := \boldsymbol{u}_l(t_n)$ for all $n\in\overline{\mathcal{N}}_\tau$. Invoking Taylor expansions, we have for all $n\in\mathcal{N}_\tau$,

$$\partial_t\boldsymbol{u}_2(t_n) = \frac{\boldsymbol{u}_2^n - \boldsymbol{u}_2^{n-1}}{\tau} + \mathcal{O}(\tau), \tag{75.13}$$

so that using $\partial_t\boldsymbol{u}_1(t_n) = \frac{\boldsymbol{u}_1^n-\boldsymbol{u}_1^{n-1}}{\tau} + \frac{\tau}{2}\partial_t\boldsymbol{u}_2(t_n) + \mathcal{O}(\tau^2)$ yields

$$\partial_t\boldsymbol{u}_1(t_n) = \frac{\boldsymbol{u}_1^n - \boldsymbol{u}_1^{n-1}}{\tau} + \frac{\tau}{2}\frac{\boldsymbol{u}_2^n - \boldsymbol{u}_2^{n-1}}{\tau} + \mathcal{O}(\tau^2), \tag{75.14}$$

and using $\partial_t\boldsymbol{u}_0(t_n) = \frac{\boldsymbol{u}_0^n-\boldsymbol{u}_0^{n-1}}{\tau} + \frac{\tau}{2}\partial_t\boldsymbol{u}_1(t_n) - \frac{\tau^2}{6}\partial_t\boldsymbol{u}_2(t_n) + \mathcal{O}(\tau^3)$ yields

$$\partial_t\boldsymbol{u}_0(t_n) = \frac{\boldsymbol{u}_0^n - \boldsymbol{u}_0^{n-1}}{\tau} + \frac{\tau}{2}\frac{\boldsymbol{u}_1^n - \boldsymbol{u}_1^{n-1}}{\tau} + \frac{\tau^2}{12}\frac{\boldsymbol{u}_2^n - \boldsymbol{u}_2^{n-1}}{\tau} + O(\tau^3). \tag{75.15}$$

We now explain the bootstrap argument to improve the accuracy. Assume that we have at hand a function r that is an $\mathcal{O}(\epsilon^l)$ approximation of $\partial_t p$ for some integer $l\in\mathbb{N}$, and assume that both p and r are smooth functions of time. Then we consider the following perturbed problem:

$$\begin{cases} \partial_t\boldsymbol{w} - \nabla\cdot\mathbb{s}(\boldsymbol{w}) + \nabla s = \boldsymbol{f}, & \boldsymbol{w}(0) = \boldsymbol{u}_0, \quad \boldsymbol{w}_{|\partial D} = \boldsymbol{0}, \\ \epsilon\partial_t s + \nabla\cdot\boldsymbol{w} = \epsilon r, & s(0) = p(0). \end{cases} \tag{75.16}$$

Owing to the stability result from Lemma 75.1, we expect that the pair $(\boldsymbol{w},s)$ is an $\mathcal{O}(\epsilon^{l+1})$ approximation of the solution $(\boldsymbol{u},p)$, i.e., the accuracy has been increased by one order. The following result formalizes this argument.

Proposition 75.6 (Bootstrapping). *Let $(\boldsymbol{u},p)$ solve (75.1), let $(\boldsymbol{w},s)$ solve (75.16), and assume that $r\in H^2(J;L^2_*(D))$. Let $\boldsymbol{e}:=\boldsymbol{u}-\boldsymbol{w}$ and $\delta:=p-s$. There is c s.t. for all $\epsilon>0$,*

$$\|\boldsymbol{e}\|^2_{L^\infty(J;\boldsymbol{L}^2)}+\epsilon\|\delta\|^2_{L^\infty(J;L^2)}+\mu\|\boldsymbol{e}\|^2_{L^2(J;\boldsymbol{V})}\le c\,\mu\epsilon^2\|\partial_t p-r\|^2_{H^1(J;L^2)},$$
$$\|\delta\|^2_{L^2(J;L^2)}\le c\left(\mu\rho^2\epsilon\|\partial_t p(0)-r(0)\|^2_{L^2}+\mu^2\epsilon^2\|\partial_t p-r\|^2_{H^2(J;L^2)}\right).$$

Proof. By the linearity of the time-dependent Stokes problem, we infer that

$$\begin{cases}\partial_t\boldsymbol{e}-\nabla\cdot\mathbb{s}(\boldsymbol{e})+\nabla\delta=\boldsymbol{0},\quad \boldsymbol{e}(0)=\boldsymbol{0},\quad \boldsymbol{e}_{|\partial D}=\boldsymbol{0},\\ \epsilon\partial_t\delta+\nabla\cdot\boldsymbol{e}=\epsilon(\partial_t p-r),\quad \delta(0)=0.\end{cases}$$

We apply Lemma 75.1 with $\boldsymbol{v}_0:=\boldsymbol{0}$, $q_0:=0$, $\boldsymbol{k}:=\boldsymbol{0}$, and $g:=\epsilon(\partial_t p-r)$. The estimate (75.4) leads to the bound on $\|\boldsymbol{e}\|^2_{L^\infty(J;\boldsymbol{L}^2)}+\epsilon\|\delta\|^2_{L^\infty(J;L^2)}+\mu\|\boldsymbol{e}\|^2_{L^2(J;\boldsymbol{V})}$. Moreover, since $\partial_t\boldsymbol{e}(0)=\boldsymbol{0}$, we can also invoke the estimate (75.5) to bound $\|\delta\|^2_{L^2(J;L^2)}$ since $\partial_t\delta(0)=\epsilon^{-1}g(0)=\partial_t p(0)-r(0)$. □

We are now in measure to construct a third-order version of the artificial compressibility method introduced in §75.2. Let $\boldsymbol{u}_l$, p_l, $\boldsymbol{f}_l$ be the l-th partial derivative of $\boldsymbol{u}$, p, $\boldsymbol{f}$ with respect to t, i.e., $\boldsymbol{u}_l:=\partial_t^l\boldsymbol{u}$, $p_l:=\partial_t^l p$, $\boldsymbol{f}_l:=\partial_t^l\boldsymbol{f}$, for all $l\in\{0,1,2\}$. Taking time derivatives of the time-dependent Stokes equations (75.1) and, for the time being, forgetting about the initial conditions on $(\boldsymbol{u}_l,p_l)$ for all $l\in\{0,1,2\}$, we have

$$\begin{cases}\partial_t\boldsymbol{u}_2-\nabla\cdot\mathbb{s}(\boldsymbol{u}_2)+\nabla p_2=\boldsymbol{f}_2,\quad \boldsymbol{u}_{2|\partial D}=\boldsymbol{0},\\ \nabla\cdot\boldsymbol{u}_2=0,\end{cases}\tag{75.17a}$$

$$\begin{cases}\partial_t\boldsymbol{u}_1-\nabla\cdot\mathbb{s}(\boldsymbol{u}_1)+\nabla p_1=\boldsymbol{f}_1,\quad \boldsymbol{u}_{1|\partial D}=\boldsymbol{0},\\ \nabla\cdot\boldsymbol{u}_1=0,\end{cases}\tag{75.17b}$$

$$\begin{cases}\partial_t\boldsymbol{u}_0-\nabla\cdot\mathbb{s}(\boldsymbol{u}_0)+\nabla p_0=\boldsymbol{f}_0,\quad \boldsymbol{u}_{0|\partial D}=\boldsymbol{0},\\ \nabla\cdot\boldsymbol{u}_0=0.\end{cases}\tag{75.17c}$$

Let us first apply the first-order artificial compressibility method to the pair $(\boldsymbol{u}_2,p_2)$, i.e., we replace $\nabla\cdot\boldsymbol{u}_2=0$ by $\epsilon\partial_t p_2+\nabla\cdot\boldsymbol{u}_2=0$ and we approximate (75.17a) as follows: For all $n\in\mathcal{N}_\tau$,

$$\begin{cases}\dfrac{1}{\tau}(\boldsymbol{u}_2^n-\boldsymbol{u}_2^{n-1})-\nabla\cdot\mathbb{s}(\boldsymbol{u}_2^n)+\nabla p_2^n=\boldsymbol{f}_2^n,\quad \boldsymbol{u}_{2|\partial D}^n=\boldsymbol{0},\\ \dfrac{\epsilon}{\tau}(p_2^n-p_2^{n-1})+\nabla\cdot\boldsymbol{u}_2^n=0,\end{cases}\tag{75.18}$$

where we have set $\boldsymbol{f}_l^n:=\boldsymbol{f}_l(t_n)$ for all $n\in\mathcal{N}_\tau$ and all $l\in\{0,1,2\}$. Proposition 75.3 shows that the sequence $(\boldsymbol{u}_{2,\tau},p_{2,\tau}):=(\boldsymbol{u}_2^n,p_2^n)_{n\in\mathcal{N}_\tau}$ is an $\mathcal{O}(\tau)$ approximation of $(\boldsymbol{u}_2,p_2)$ (at least informally and if $(\boldsymbol{u}_2(0),p_2(0))$ is known).

Let us now consider the pair $(\boldsymbol{u}_1,p_1)$. Using (75.14) to approximate $\partial_t\boldsymbol{u}_1(t_n)$ in (75.17b) gives a second-order accurate approximation. Next, we replace $\nabla\cdot\boldsymbol{u}_1=0$ by $\epsilon\partial_t p_1+\nabla\cdot\boldsymbol{u}_1=\epsilon r$, where r is some $\mathcal{O}(\tau)$ approximation of $\partial_t p_1$. But recalling that the purpose of p_2 is precisely to approximate $\partial_t p_1$, we are going to substitute r by p_2. Then replacing $\partial_t p_1(t_n)$ by the first-order approximation $\frac{1}{\tau}(p_1^n-p_1^{n-1})$ and putting everything together gives the following time discretization of (75.17b): For all $n\in\mathcal{N}_\tau$,

$$\begin{cases}\dfrac{1}{\tau}(\boldsymbol{u}_1^n-\boldsymbol{u}_1^{n-1})-\nabla\cdot\mathbb{s}(\boldsymbol{u}_1^n)+\nabla p_1^n=\widetilde{\boldsymbol{f}}_1^n,\quad \boldsymbol{u}_{1|\partial D}^n=\boldsymbol{0},\\ \dfrac{\epsilon}{\tau}(p_1^n-p_1^{n-1})+\nabla\cdot\boldsymbol{u}_1^n=\epsilon p_2^n,\end{cases}\tag{75.19}$$

where $\widetilde{\boldsymbol{f}}_1^n := \boldsymbol{f}_1^n - \frac{\tau}{2}\delta_\tau \boldsymbol{u}_2^n$ and $\delta_\tau \boldsymbol{v}^n := \frac{\boldsymbol{v}^n - \boldsymbol{v}^{n-1}}{\tau}$ for any sequence $(\boldsymbol{v}^n)_{n\in\mathcal{N}_\tau}$. From Proposition 75.6, we expect (75.19) to give an $\mathcal{O}(\epsilon\tau) = \mathcal{O}(\tau^2)$ approximation of the pair $(\boldsymbol{u}_1, p_1)$ (at least informally and if $(\boldsymbol{u}_1(0), p_1(0))$ is known).

Let us finally take the reasoning one step further by considering the pair $(\boldsymbol{u}_0, p_0)$. Using (75.15) to approximate $\partial_t \boldsymbol{u}_0(t_n)$ in (75.17c) gives a third-order accurate approximation. Next we replace $\nabla\cdot\boldsymbol{u}_0 = 0$ by $\epsilon\partial_t p_0 + \nabla\cdot\boldsymbol{u}_0 = \epsilon r$, where r is some $\mathcal{O}(\tau^2)$ approximation of $\partial_t p_2$. But recalling that the purpose of p_1 is precisely to approximate $\partial_t p_0$, we are going to substitute r by p_1. Then replacing $\partial_t p_0(t_n)$ by the second-order approximation $\frac{1}{\tau}(p_0^n - p_0^{n-1}) + \frac{1}{2}\tau p_2^n$ (this follows from the Taylor expansion $p_0^{n-1} = p_0^n - \tau\partial_t p_0(t_n) + \frac{1}{2}\tau^2\partial_{tt} p_0(t_n) + \mathcal{O}(\tau^3)$) and putting everything together gives the following time discretization of (75.17c): For all $n \in \mathcal{N}_\tau$,

$$\begin{cases} \dfrac{1}{\tau}(\boldsymbol{u}_0^n - \boldsymbol{u}_0^{n-1}) - \nabla\cdot\mathbb{s}(\boldsymbol{u}_0^n) + \nabla p_0^n = \widetilde{\boldsymbol{f}}_0^n, \quad \boldsymbol{u}_{0|\partial D}^n = \boldsymbol{0}, \\ \dfrac{\epsilon}{\tau}(p_0^n - p_0^{n-1}) + \nabla\cdot\boldsymbol{u}_0^n = \epsilon p_1^n - \dfrac{1}{2}\epsilon\tau p_2^n, \end{cases} \tag{75.20}$$

with the shorthand notation $\widetilde{\boldsymbol{f}}_0^n := \boldsymbol{f}_0^n - \frac{\tau}{2}\delta_\tau \boldsymbol{u}_1^n - \frac{\tau^2}{12}\delta_\tau \boldsymbol{u}_2^n$.

After uncoupling the velocity and the pressure in (75.18)-(75.19)-(75.20), the final form of the algorithm proceeds as follows: For all $n \in \mathcal{N}_\tau$,

$$\begin{cases} \boldsymbol{u}_2^n - \tau\big(\nabla\cdot\mathbb{s}(\boldsymbol{u}_2^n) + \lambda\nabla\nabla\cdot\boldsymbol{u}_2^n\big) = \boldsymbol{g}_2^n, \quad \boldsymbol{u}_{2|\partial D}^n = \boldsymbol{0}, \\ p_2^n = p_2^{n-1} - \lambda\nabla\cdot\boldsymbol{u}_2^n, \end{cases} \tag{75.21a}$$

$$\begin{cases} \boldsymbol{u}_1^n - \tau\big(\nabla\cdot\mathbb{s}(\boldsymbol{u}_1^n) + \lambda\nabla\nabla\cdot\boldsymbol{u}_1^n\big) = \boldsymbol{g}_1^n, \quad \boldsymbol{u}_{1|\partial D}^n = \boldsymbol{0}, \\ p_1^n = p_1^{n-1} + \tau p_2^n - \lambda\nabla\cdot\boldsymbol{u}_1^n, \end{cases} \tag{75.21b}$$

$$\begin{cases} \boldsymbol{u}_0^n - \tau\big(\nabla\cdot\mathbb{s}(\boldsymbol{u}_0^n) + \lambda\nabla\nabla\cdot\boldsymbol{u}_0^n\big) = \boldsymbol{g}_0^n, \quad \boldsymbol{u}_{0|\partial D}^n = \boldsymbol{0}, \\ p_0^n = p_0^{n-1} + \tau p_1^n - \frac{1}{2}\tau^2 p_2^n - \lambda\nabla\cdot\boldsymbol{u}_0^n, \end{cases} \tag{75.21c}$$

with

$$\begin{aligned} \boldsymbol{g}_2^n &:= \boldsymbol{u}_2^{n-1} + \tau\big(\boldsymbol{f}_2^n - \nabla p_2^{n-1}\big), \\ \boldsymbol{g}_1^n &:= \boldsymbol{u}_1^{n-1} + \tau\Big(\boldsymbol{f}_1^n - \nabla\big(p_1^{n-1} + \tau p_2^n\big) - \frac{1}{2}\delta_\tau \boldsymbol{u}_2^n\Big), \\ \boldsymbol{g}_0^n &:= \boldsymbol{u}_0^{n-1} + \tau\Big(\boldsymbol{f}_0^n - \nabla\big(p_0^{n-1} + \tau p_1^n - \frac{\tau^2}{2} p_2^n\big) - \frac{\tau}{2}\delta_\tau \boldsymbol{u}_1^n - \frac{\tau^2}{12}\delta_\tau \boldsymbol{u}_2^n\Big). \end{aligned}$$

This shows that each step of the third-order artificial compressibility method requires to solve three implicit parabolic time steps.

Remark 75.7 (Initialization). The initialization of the scheme (75.21a)–(75.21c) requires the specification of $(\boldsymbol{u}_l(0), p_l(0))$ for all $l \in \{0, 1, 2\}$. This is the price to pay for replacing $\nabla\cdot\boldsymbol{u} = 0$ by $\epsilon\partial_t p + \nabla\cdot\boldsymbol{u} = 0$. The initialization of $p_0(0)$ is discussed in Remark 74.4. A third-order initialization strategy is proposed in Exercise 75.4. Notice that the initialization is trivial if the initial state is rest and the source term starts very smoothly from zero, i.e., $\boldsymbol{u}_0 = \boldsymbol{0}$, $\boldsymbol{f}(0) = \boldsymbol{0}$, $\partial_t \boldsymbol{f}(0) = \boldsymbol{0}$, and $\partial_{tt}\boldsymbol{f}(0) = \boldsymbol{0}$. □

Remark 75.8 (Navier–Stokes). A nonlinear version of the above scheme has been proposed in Guermond and Minev [152, Eq. (3.12)-(3.14)] to solve the Navier–Stokes equations. □

75.4 Finite element implementation

We now give some details on how the above algorithm can be implemented with mixed finite elements. Let $\mathcal{M}$ be the mass matrix for the velocity. Let $\mathcal{N}$ be the mass matrix for the pressure, or its lumped version, or any diagonal matrix with entries scaling like those of the mass matrix (say, for each row i, take the volume of the support of the i-th pressure shape function). Let $\mathcal{A}$ be the stiffness matrix associated with the operator $-\nabla\cdot\mathbb{s}(\cdot)$. Similarly, we denote by $\mathcal{B}$ the matrix associated with the divergence operator $\nabla\cdot$. Then $\mathcal{B}^{\mathsf{T}}$ is the matrix associated with the negative of the gradient operator.

At every time step the fully discrete versions of the systems (75.18)-(75.19)-(75.20) require solving linear systems of the form $(\frac{1}{\tau}\mathcal{M}+\mathcal{A})\mathsf{U} = \mathsf{F} + \lambda\mathcal{B}^{\mathsf{T}}\mathsf{P}$ and $\mathcal{N}\mathsf{P} = \mathcal{N}\mathsf{Q} - \lambda\mathcal{B}\mathsf{U}$. Notice here that the exact matrix version of $p = q - \lambda\nabla\cdot\boldsymbol{u}$ induced by the Galerkin formulation implies that $\mathcal{N}$ is the pressure mass matrix. But this constraint can be relaxed since, without loss of accuracy, instead of approximating $\epsilon\partial_t p + \nabla\cdot\boldsymbol{u} = 0$, we could also approximate the perturbation $\epsilon\partial_t L(p) + \nabla\cdot\boldsymbol{u} = 0$, where L is any perturbation of the identity operator in the pressure space. As said above, instead of using the consistent mass matrix for the pressure, one does not lose the properties of the scheme by using either the lumped mass matrix or any appropriately scaled diagonal matrix. In conclusion, one eliminates the pressure in the velocity equation by using $\mathsf{P} = \mathsf{Q} - \lambda\mathcal{N}^{-1}\mathcal{B}\mathsf{U}$, and one obtains $(\frac{1}{\tau}\mathcal{M} + \mathcal{A} + \lambda\mathcal{B}^{\mathsf{T}}\mathcal{N}^{-1}\mathcal{B})\mathsf{U} = \mathsf{F} + \lambda\mathcal{B}^{\mathsf{T}}\mathsf{Q}$. We insist again that $\mathcal{N}$ need not be the consistent pressure mass matrix. Actually, we recommend to use either the lumped mass matrix or any appropriately scaled diagonal matrix. We refer the reader to [152] for additional details on this technique. The above method has been tested in [152] and has been shown numerically to deliver third-order accuracy in time on the velocity and the pressure in all the relevant norms. Fourth order and higher orders can be obtained by using the appropriate Taylor expansions.

Exercises

Exercise 75.1 (Lemma 75.1). (i) Prove (75.4). (*Hint*: test the momentum equation with $\boldsymbol{v}$ and the mass equation with q, use Lemma 53.9 to bound $(q,g)_{L^2}$, integrate in time from 0 to t for all $t \in J$, and integrate by parts in time.) (ii) Prove (75.5). (*Hint*: use the inf-sup condition on the bilinear form b together with the bounds derived in Step (i).)

Exercise 75.2 (Lemma 75.2). (i) Let $\delta_\tau\boldsymbol{k}^n := \frac{\boldsymbol{k}^n-\boldsymbol{k}^{n-1}}{\tau}$ and $\delta_\tau g^n := \frac{g^n-g^{n-1}}{\tau}$ for all $n \in \mathcal{N}_\tau$. Prove that $\|\delta_\tau\boldsymbol{k}_\tau\|_{\ell^2(J_*;\boldsymbol{L}^2)} \le \|\partial_t\boldsymbol{k}\|_{L^2(J_*;\boldsymbol{L}^2)}$. Let $\Gamma(t) := \frac{1}{\tau}\int_{t-\tau}^t \partial_\xi g(\xi)\,\mathrm{d}\xi$ for all $t \in J_*$. Prove that $\partial_t\Gamma(t) = \frac{1}{\tau}\int_{t-\tau}^t \partial_{\xi\xi} g(\xi)\,\mathrm{d}\xi$ for all $t \in J_*$ and that $\|\partial_t\Gamma\|_{L^2(J_{**},L^2)} \le \|\partial_{\xi\xi} g\|_{L^2(J_*;L^2)}$. (*Hint*: use the Cauchy–Schwarz inequality and Fubini's theorem.) (ii) Derive the system satisfied by the time sequences $\delta_\tau\boldsymbol{u}_\tau := (\frac{\boldsymbol{u}^n-\boldsymbol{u}^{n-1}}{\tau})_{n\in\mathcal{N}_\tau}$ and $\delta_\tau p_\tau := (\frac{p^n-p^{n-1}}{\tau})_{n\in\mathcal{N}_\tau}$. (iii) Prove the estimate (75.10). (*Hint*: use the inf-sup condition on the bilinear form b and bound $\delta_\tau\boldsymbol{u}_\tau$ by adapting the proof of (75.9).)

Exercise 75.3 (Proposition 75.3). The goal of this exercise is to prove Proposition 75.3. (i) Let $\boldsymbol{e}_\tau := \boldsymbol{u}_\tau - \boldsymbol{\pi}_\tau(\boldsymbol{u})$ and $r_\tau := p_\tau - \pi_\tau(p)$. Let $\boldsymbol{\psi}(t) := \frac{1}{\tau}\int_{t-\tau}^t(\xi - t + \tau)\partial_{\xi\xi}\boldsymbol{u}\,\mathrm{d}\xi$ and $\phi(t) :=$

$-\frac{1}{\lambda}\int_{t-\tau}^{t}\partial_\xi p\,\mathrm{d}\xi$ for all $t\in J_*$. Show that

$$\begin{cases}\dfrac{1}{\tau}(\boldsymbol{e}^n-\boldsymbol{e}^{n-1})-\nabla\cdot\mathbb{s}(\boldsymbol{e}^n)+\nabla r^n=\boldsymbol{\psi}^n,\quad \boldsymbol{e}^n_{|\partial D}=\boldsymbol{0},\\ \dfrac{1}{\lambda}(r^n-r^{n-1})+\nabla\cdot\boldsymbol{e}^n=\phi^n.\end{cases}$$

(ii) Prove the estimates (75.11) and (75.12). (*Hint*: use Lemma 75.2.)

Exercise 75.4 (Initialization). Let $\boldsymbol{u}_0$ be the initial velocity, and assume that $p(0)$ is given. Let $t_1:=\tau$. Using the first-order artificial compressibility algorithm (75.6) and Richardson's extrapolation, propose a technique to estimate $(\partial_{tt}\boldsymbol{u}(t_1),\partial_{tt}p(t_1))$ with $\mathcal{O}(\tau)$ accuracy, $(\partial_t\boldsymbol{u}(t_1),\partial_t p(t_1))$ with $\mathcal{O}(\tau^2)$ accuracy, and $(\boldsymbol{u}(t_1),p(t_1))$ with $\mathcal{O}(\tau^3)$ accuracy. (*Hint*: estimate $(\boldsymbol{u},p)$ at the times t_1 and $t_2:=2\tau$ by using (75.6) with the time steps $\frac{\tau}{3}$, $\frac{\tau}{2}$, and τ, keeping λ fixed. Conclude by using finite differences centered at $t_1:=\tau$.)

Chapter 76

Well-posedness and space semi-discretization

The three chapters composing Part XV deal with the approximation of time-dependent Friedrichs' systems and more generally systems of first-order PDEs. We study the approximation of these systems by using implicit and explicit time-stepping techniques combined with stabilized finite elements. The prototypical example we have in mind is the linear transport equation where one seeks a space-time function $u : D\times J \to \mathbb{R}$ s.t.

$$\partial_t u + \boldsymbol{\beta}\cdot\nabla u = f \quad \text{in } D\times J, \tag{76.1}$$

where D is a Lipschitz domain in $\mathbb{R}^d$, $J := (0,T)$ is the time interval with $T > 0$, $\boldsymbol{\beta} : D \to \mathbb{R}^d$ is the transport velocity, and $f : D\times J \to \mathbb{R}$ is the source term. More generally, u and f are $\mathbb{C}^m$-valued, $m \geq 1$, and the generic form of the evolution problem (76.1) is

$$\partial_t u + A(u) = f \quad \text{in } D\times J, \tag{76.2}$$

where A is a Friedrichs' operator, i.e., the problem $A(u) = f$ is one of the symmetric positive systems of first-order linear PDEs introduced in Chapter 56. In this chapter, we first derive a functional setting for (76.2) and establish its well-posedness. For simplicity, we assume that the differential operator in space is time-independent, e.g., the transport velocity in (76.1) is time-independent. Then we construct a space semi-discretization of the problem using stabilized finite elements. We focus on the fluctuation-based stabilization techniques introduced in Chapters 58-59. Implicit and explicit time discretization techniques are investigated in the next chapter.

76.1 Maximal monotone operators

The notion of maximal monotone operators provides a suitable functional setting to formulate noncoercive time-dependent problems. Let L be a separable (real or complex) Hilbert space. We identify L with L'. Let V_0 be a proper subspace of L and let $A : V_0 \to L$ be a linear operator. The setting we have in mind is

$$V_0 \subsetneq V := D(A) \subsetneq L, \tag{76.3}$$

where $D(A) := \{v \in L \mid A(v) \in L\}$ is called the graph (or the domain) of A. In general, the operator A is unbounded on L, i.e., V is a proper subspace of L (if A is bounded on L, then

$V = L$). The purpose of V_0 is to enforce appropriate boundary conditions. More precisely, we assume that V_0 is such that the restriction of A to V_0 is maximal montone in the sense defined below. We denote by $I_X : X \to X$ the identity operator for any subspace $X \subset L$.

Definition 76.1 (Maximal monotone operator). *The operator $A : V_0 \to L$ is said to be* monotone *if*

$$\Re\big((A(v), v)_L\big) \geq 0, \qquad \forall v \in V_0, \tag{76.4}$$

and it is said to be maximal monotone *if in addition there exists a real number $\tau_0 > 0$ such that $I_{V_0} + \tau_0 A : V_0 \to L$ is surjective:*

$$\forall f \in L, \quad \exists v \in V_0 \quad \text{s.t.} \quad v + \tau_0 A(v) = f. \tag{76.5}$$

Given a source term $f \in C^1(\overline{J}; L)$ and an initial condition $u_0 \in V_0$, we consider the following model problem:

$$\begin{cases} \text{Find } u \in C^1(\overline{J}; L) \cap C^0(\overline{J}; V_0) \text{ s.t. } u(0) = u_0 \text{ and} \\ \partial_t u(t) + A(u(t)) = f(t), \qquad \forall t \in \overline{J}. \end{cases} \tag{76.6}$$

Notice that in this setting the time derivative is defined in the strong sense.

Remark 76.2 (Time scale). The formulation of the model problem (76.6) shows that the operator A has the same dimension as the reciprocal of a time. Therefore, the real number τ_0 in (76.5) is a time scale. □

Example 76.3 (Transport operator). Let us set $A(v) := \boldsymbol{\beta}\cdot\nabla v$ with $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(D)$ and assume for simplicity that $\nabla\cdot\boldsymbol{\beta} = 0$. Let $L := L^2(D)$ and $V := \{v \in L^2(D) \mid \boldsymbol{\beta}\cdot\nabla v \in L^2(D)\}$. Notice that V is a proper subspace of L, i.e., A is unbounded on L. Let the inflow and outflow parts of the boundary be $\partial D^\pm := \{\boldsymbol{x} \in \partial D \mid \pm(\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x}) > 0\}$ and assume that ∂D^- and ∂D^+ are well-separated. Then, as discussed in Example 56.13, the trace operator $\gamma : C^0(\overline{D}) \to C^0(\partial D)$ s.t. $\gamma(v) = v_{|\partial D}$ can be extended to a bounded linear operator from V to $L^2_{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}(\partial D; \mathbb{R})$, where the subscript $|\boldsymbol{\beta}\cdot\boldsymbol{n}|$ means that the measure $\mathrm{d}s$ is replaced by $|\boldsymbol{\beta}\cdot\boldsymbol{n}|\,\mathrm{d}s$. Owing to the integration by parts formula $(\boldsymbol{\beta}\cdot\nabla v, w)_L + (v, \boldsymbol{\beta}\cdot\nabla w)_L = ((\boldsymbol{\beta}\cdot\boldsymbol{n})\gamma(v), \gamma(w))_{L^2(\partial D)}$ for all $v, w \in V$, we have

$$(A(v), v)_L = \frac{1}{2}\|\gamma(v)\|^2_{L^2(|\boldsymbol{\beta}\cdot\boldsymbol{n}|;\partial D^+)} - \frac{1}{2}\|\gamma(v)\|^2_{L^2(|\boldsymbol{\beta}\cdot\boldsymbol{n}|;\partial D^-)},$$

which shows that A is not monotone on V but is monotone on the proper subspace $V_0 := \{v \in V \mid \gamma(v)_{|\partial D^-} = 0\}$. Hence, enforcing the condition $\gamma(v) = 0$ at the inflow boundary ∂D^- yields the monotonicity property (76.4). Moreover, the well-posedness theory for Friedrichs' systems (see Theorem 56.9) shows that A is maximal monotone for any real number $\tau_0 > 0$. □

Lemma 76.4 (Density, Hilbert space). *Let $A : V_0 \to L$ be a maximal monotone operator. The following properties hold true:*

(i) *V_0 is dense in L.*

(ii) *$I_{V_0} + \tau_0 A : V_0 \to L$ is an isomorphism and $\|(I_{V_0} + \tau_0 A)^{-1}\|_{\mathcal{L}(L;L)} \leq 1$.*

(iii) *Equipped with the graph norm $\|v\|_V^2 := \|v\|_L^2 + \tau_0^2\|A(v)\|_L^2$ and the associated inner product $(v, w)_L + \tau_0^2(A(v), A(w))_L$, V_0 is a Hilbert space.*

Proof. (i) Let us apply Corollary C.15 which gives a characterization for density. Let $f \in L \equiv L'$ be such that $(f, v)_L = 0$ for all $v \in V_0$. Since $I_{V_0} + \tau_0 A : V_0 \to L$ is surjective owing to the

maximality property, there is $v_0 \in V_0$ so that $v_0 + \tau_0 A(v_0) = f$. The monotonicity property implies that

$$\|v_0\|_L^2 \le \|v_0\|_L^2 + \tau_0 \Re\big((A(v_0), v_0)_L\big) = \Re\big((v_0 + \tau_0 A(v_0), v_0)_L\big) = \Re\big((f, v_0)_L\big) = 0.$$

Hence, $v_0 = 0$, i.e., $f = 0$. This shows that V_0 is dense in L.
(ii) Let us set $B := I_{V_0} + \tau_0 A : V_0 \to L$. Maximality means that B is surjective. Monotonicity implies that B is also injective since $B(v) = 0$ implies that $0 = \Re((B(v), v)_L) = \|v\|_L^2 + \tau_0 \Re((A(v), v)_L) \ge \|v\|_L^2$, so that $v = 0$. Hence, for all $f \in L$, there exists a unique $v := B^{-1}(f) \in V_0$ so that $B(v) = v + \tau_0 A(v) = f$. Since $\|v\|_L^2 \le \|v\|_L^2 + \tau_0 \Re((A(v), v)_L) = \Re((f, v)_L) \le \|f\|_L \|v\|_L$, we have $\|v\|_L \le \|f\|_L$. This shows that $\|B^{-1}\|_{\mathcal{L}(L;L)} \le 1$.
(iii) Let $(v_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in V_0, i.e., the sequence is Cauchy for the graph norm $\big(\|v_n\|_L^2 + \tau_0^2 \|A(v_n)\|_L^2\big)^{\frac12}$. This implies that both $(v_n)_{n\in\mathbb{N}}$ and $(A(v_n))_{n\in\mathbb{N}}$ are Cauchy sequences in L. Hence, there are $v \in L$ and $f \in L$ so that $v_n \to v$ and $A(v_n) \to f$ as $n \to \infty$. Using the boundedness of $(I_{V_0} + \tau_0 A)^{-1} : L \to V_0 \subset L$ and since $v_n \in V_0$, we have

$$v \leftarrow v_n = (I_{V_0} + \tau_0 A)^{-1}(v_n + \tau_0 A(v_n)) \to (I_{V_0} + \tau_0 A)^{-1}(v + \tau_0 f).$$

Hence, $v = (I_{V_0} + \tau_0 A)^{-1}(v + \tau_0 f)$. This shows that v is in the range of $(I_{V_0} + \tau_0 A)^{-1}$, i.e., $v \in V_0$, and $A(v) = f$, i.e., $\|v_n - v\|_V \to 0$. □

Corollary 76.5 (Bijectivity). *Let $A : V_0 \to L$ be a maximal monotone operator. For any real number $\tau > 0$, the linear operator $I_{V_0} + \tau A : V_0 \to L$ is bijective, and we have $\|(I_{V_0} + \tau A)^{-1}\|_{\mathcal{L}(L;L)} \le 1$.*

Proof. (1) Since the norms $(\|v\|_L^2 + \tau_0^2 \|A(v)\|_L^2)^{\frac12}$ and $(\|v\|_L^2 + \tau^2 \|A(v)\|_L^2)^{\frac12}$ are equivalent for all $\tau_0, \tau > 0$, and since V_0 is a Hilbert space when equipped with the former inner product by Lemma 76.4(iii), it is also a Hilbert space when equipped with the latter inner product. Consider now the bilinear form $a(v, w) := (v + \tau A(v), w)_L$ defined on $V_0 \times L$. We need to verify the two conditions of the BNB theorem (Theorem 25.9).
(2) Let us first prove (BNB1). Let $v \in V_0$ and set $S(v) := \sup_{w\in L} \frac{|a(v,w)|}{\|w\|_L}$. Taking $w := v \in V_0 \subset L$ in the definition of $S(v)$ and invoking the monotonicity property of A, we infer that

$$S(v) \ge \frac{|a(v, v)|}{\|v\|_L} \ge \frac{\Re\big(a(v, v)\big)}{\|v\|_L} \ge \frac{\|v\|_L^2}{\|v\|_L} = \|v\|_L.$$

Moreover, taking $w := A(v) \in L$ in the definition of $S(v)$ and since

$$\begin{aligned} |a(v, A(v))| &= |(v, A(v))_L + \tau \|A(v)\|_L^2| \\ &\ge \Re\big((v, A(v))_L + \tau \|A(v)\|_L^2\big) \ge \tau \|A(v)\|_L^2, \end{aligned}$$

owing again to the monotonicity property of A, we infer that

$$S(v) \ge \frac{|a(v, A(v))|}{\|A(v)\|_L} \ge \tau \|A(v)\|_L.$$

Combining the two above bounds leads to $\sqrt{2} S(v) \ge \|v\|_V$. This shows that (BNB1) holds true with the constant $\frac{1}{\sqrt{2}}$.
(3) We now prove (BNB2). Let $w \in L$ and assume that $a(v, w) = 0$ for all $v \in V_0$. Since

$I_{V_0} + \tau_0 A : V_0 \to L$ is bijective by Lemma 76.4(ii), there is $v_0 \in V_0$ such that $v_0 + \tau_0 A(v_0) = w$. Using the monotonicity of A, we obtain

$$\begin{aligned} 0 &= \Re\big(a(v_0, w)\big) = \Re\big((v_0 + \tau A(v_0), v_0 + \tau_0 A(v_0))_L\big) \\ &= \|v_0\|_L^2 + (\tau + \tau_0)\Re\big((A(v_0), v_0)_L\big) + \tau\tau_0 \|A(v_0)\|_L^2 \geq \|v_0\|_L^2. \end{aligned}$$

Hence, $v_0 = 0$, which implies $w = 0$ and proves (BNB2). We establish that $\|(I_{V_0} + \tau A)^{-1}\|_{\mathcal{L}(L;L)} \leq 1$ by proceeding as in the proof of Lemma 76.4(ii). □

Remark 76.6 (Literature). The reader is referred to Showalter [257, p. 22], Yosida [289, p. 246], Brezis [52, Prop. 7.1] for more details on maximal monotone operators. An interesting physics-oriented extension of the theory is presented in Picard [238]. □

76.2 Well-posedness

The setting of maximal monotone operators is useful to derive an existence and uniqueness result for the time evolution problem (76.6). The main interest of this setting lies in the fact that the study of the evolution problem reduces to the study of the properties of the operator $A : V_0 \to L$. The price to pay to use this setting is that the operator A is time-independent. The main result of this section is the following.

Theorem 76.7 (Hille–Yosida). *Let $A : V_0 \to L$ be a maximal monotone operator. For all $f \in C^1(\overline{J}; L)$ and all $u_0 \in V_0$, there exists a unique $u \in C^1(\overline{J}; L) \cap C^0(\overline{J}; V_0)$ solving* (76.6). *Moreover, the following a priori estimate holds true: For all $t \in \overline{J}$,*

$$\|u(t)\|_L \leq e^{\frac{t}{2T}}(tT)^{\frac{1}{2}}\|f\|_{C^0([0,t];L)} + \|u_0\|_L. \tag{76.7}$$

In particular, for $t = T$ we have $\|u(T)\|_L \leq e^{\frac{1}{2}}T\|f\|_{C^0(\overline{J};L)} + \|u_0\|_L$.

Proof. For the existence and uniqueness result, we refer the reader to Yosida [289, p. 248] or Brezis [52, Thm. 7.4]. Notice that one cannot invoke Lions' theorem (Theorem 65.9) since the bilinear form $(v, w) \mapsto (A(v), w)_L$ is not coercive on $V \times V$. Let us prove the a priori estimate (76.7). Let $u_1 \in C^1(\overline{J}; L) \cap C^0(\overline{J}; V_0)$ solve $\partial_t u_1(t) + A(u_1(t)) = 0$ for all $t \in \overline{J}$ and $u_1(0) = u_0$, and let $u_2 \in C^1(\overline{J}; L) \cap C^0(\overline{J}; V_0)$ solve $\partial_t u_2(t) + A(u_2(t)) = f$ for all $t \in \overline{J}$ and $u_2(0) = 0$. By linearity, we have $u = u_1 + u_2$, so that we are going to bound u_1 and u_2 separately. Since the equation $\partial_t u_1(t) + A(u_1(t)) = 0$ holds in $C^0(\overline{J}; L)$, we can take the L-inner product of this equation with $u_1(t)$ for all $t \in \overline{J}$ and infer that

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_1(t)\|_L^2 &= \Re\big((\partial_t u_1(t), u_1(t))_L\big) \\ &\leq \Re\big(\partial_t u_1(t) + A(u_1(t)), u_1(t))_L\big) = 0, \end{aligned}$$

where we used the monotonicity of A. Similar arguments show that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_2(t)\|_L^2 \leq \Re\big((\partial_t u_2(t) + A(u_2(t)), u_2(t))_L\big) = \Re\big((f(t), u_2(t))_L\big).$$

Integrating the above two estimates on $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_1(t)\|_L^2$ and $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u_2(t)\|_L^2$ from 0 to t for all $t \in \overline{J}$, and using the initial conditions $u_1(0) = u_0$, $u_2(0) = 0$ and the Cauchy–Schwarz inequality, we infer

that

$$\|u_1(t)\|_L^2 \le \|u_0\|_L^2, \qquad \|u_2(t)\|_L^2 \le \int_0^t 2\|f(s)\|_L\|u_2(s)\|_L \,\mathrm{d}s.$$

Invoking a modified version of Gronwall's lemma (see Lemma 76.8 below, with $\phi(t) := \|u_2(t)\|_L^2$, $a(t) := 2\|f(t)\|_L$, and $b(t) := 0$), we infer that we have for all $t \in \overline{J}$,

$$\|u_2(t)\|_L^2 \le e^{\frac{t}{T}} T\|f\|_{L^2([0,t];L)}^2 \le e^{\frac{t}{T}} tT\|f\|_{C^0([0,t];L)}^2,$$

where the last bound follows from the Cauchy–Schwarz inequality. Since $\|u(t)\|_L \le \|u_1(t)\|_L + \|u_2(t)\|_L$, this completes the proof. □

Lemma 76.8 (Gronwall's lemma). *Let $\phi \in L^1(J;\mathbb{R}_+)$. Assume that there is a function $a \in L^2(J;\mathbb{R})$ and a nondecreasing function $b \in L^1(J;\mathbb{R})$ such that $\phi(t) \le \int_0^t a(s)\phi(s)^{\frac12}\,\mathrm{d}s + b(t)$ for all $t \in J$. The following holds true for all $t \in \overline{J}$ (with the convention that $\|a\|_{L^2(0,t)} = 0$ if $t = 0$):*

$$\phi(t) \le e^{\frac{t}{T}}\big(\tfrac{T}{4}\|a\|_{L^2(0,t)}^2 + b(t)\big). \tag{76.8}$$

Proof. Since $a(s)\phi(s)^{\frac12} \le \frac{Ta(s)^2}{4} + \frac{\phi(s)}{T}$ owing to Young's inequality, we obtain

$$\phi(t) \le \frac{T}{4}\|a\|_{L^2(0,t)}^2 + b(t) + \int_0^t \frac{\phi(s)}{T}\,\mathrm{d}s.$$

We now apply Gronwall's lemma (see (65.20) from Exercise 65.3) with $\alpha(t) := \frac{T}{4}\|a\|_{L^2(0,t)}^2 + b(t)$ and $\beta(t) := \frac{1}{T}$. This yields the assertion. □

Remark 76.9 (Time growth/decay). Contrary to the parabolic setting where the influence of the initial data decays exponentially fast as t grows (see Lemma 65.11), the estimate (76.7) shows that the influence of the initial data is permanent. Moreover, the source term f may induce a linear growth of $\|u(T)\|_L$ with respect to T. This is a characteristic property of evolution PDEs without coercivity. However, if it turns out that there is $\mu_\sharp > 0$ s.t. $\Re\big((A(v),v)_L\big) \ge \mu_\sharp\|v\|_L^2$ for all $v \in V_0$ (which is a stronger property than the monotonicity property (76.4)), then the estimate (76.7) can be replaced by the following sharper estimate:

$$\|u(t)\|_{L^2}^2 \le e^{-\frac{\mu_\sharp}{2}t}\|u_0\|_L^2 + \frac{1}{\mu_\sharp}\int_0^t e^{-\mu_\sharp(t-s)}\|f(s)\|_L^2\,\mathrm{d}s, \tag{76.9}$$

for all $t \in J$; see Exercise 76.3. □

Remark 76.10 (Dimensionality and Young's inequality). Notice that blindly applying Young's inequality in the form $\alpha\beta \le \frac14\alpha^2 + \beta^2$ is questionable from the dimension point of view if the real numbers α, β do not have the same dimension. For instance, the inequality invoked in the proof of Lemma 76.8 is dimensionally consistent since the bound $\phi(t) \le \int_0^t a(s)\phi(s)^{\frac12}\,\mathrm{d}s + b(t)$ shows that the square root of the dimension of ϕ is equal to the dimension of a multiplied by a time scale. More generally, we could have invoked Young's inequality in the form $a(s)\phi(s)^{\frac12} \le \frac{\theta a(s)^2}{4} + \frac{\phi(s)}{\theta}$, where θ is any time scale, but the sharpest choice is $\theta = T$ (see Exercise 76.2). □

Remark 76.11 (Contraction semigroups). Let $f := 0$. Then Theorem 76.7 shows that the linear map $V_0 \ni u_0 \mapsto u(t) \in V_0 \subset L$ is s.t. $\|u(t)\|_L \le \|u_0\|_L$ for all $t \ge 0$. Since V_0 is dense in L, we can extend this map by density. Let us denote by $S_A(t) : L \ni u_0 \mapsto u(t) \in L$ the

bounded linear operator thus defined. This operator has the following properties: (i) For all $t \geq 0$, $S_A(t) \in \mathcal{L}(L;L)$ and $\|S_A(t)\|_{\mathcal{L}(L;L)} \leq 1$; (ii) For all $t_1, t_2 \geq 0$, $S_A(t_1+t_2) = S_A(t_1) \circ S_A(t_2)$; (iii) $\lim_{t \downarrow 0} \|S_A(t)(u_0) - u_0\|_L = 0$ for all $u_0 \in L$. Any family of operators $(R(t))_{t \in \mathbb{R}_+}$ in $\mathcal{L}(L;L)$ satisfying the above three properties is called a *contraction semigroup* of class C^0. A striking result is that for every contraction semigroup of class C^0 on L, say $(R(t))_{t \in \mathbb{R}_+}$, there exists a unique maximal monotone operator A such that $R(t) = S_A(t)$; see e.g., Yosida [289, Chap. 9] (see in particular the theorem p. 246 and the Phillips–Lumer theorem p. 250 therein) and Brezis [52, Rem. 5]. □

Remark 76.12 (Hille–Yosida vs. Lions). On the one hand the Hille–Yosida theorem is slightly more general than Lions's theorem since it does not require the bilinear form $(v,w) \mapsto (A(v),w)_L$ to be V-coercive. On the other hand the Hille–Yosida theorem is somewhat more restrictive since the time derivative is taken in strong form, and this is reflected by the relatively strong assumptions made on f and u_0. The solution considered in the Hille–Yosida theorem is called *strong solution*, whereas the one considered in Lions' theorem is called *weak solution*. It is however possible to weaken the assumptions on f and u_0 in Theorem 76.7. In particular, it is shown in Ball [22] that for all $f \in L^1(J;L)$ and all $u_0 \in L$, there exists a unique $u \in L^1(J;L)$ such that we have $\int_0^t u(s)\,\mathrm{d}s \in V_0$ and $u(t) = u_0 + A\big(\int_0^t u(s)\,\mathrm{d}s\big) + \int_0^t f(s)\,\mathrm{d}s$ for all $t \in J$. This solution is given by $u(t) = S_A(t)(u_0) + \int_0^t S_A(t-s)(f(s))\,\mathrm{d}s$. This type of solution is often called *mild solution* in the literature. □

If instead of being monotone, $A : V_0 \to L$ satisfies the weaker assumption

$$\exists \mu_\flat > 0, \quad \forall v \in V_0, \quad \Re\big((A(v),v)_L\big) \geq -\mu_\flat \|v\|_L^2, \tag{76.10}$$

then $\mu_\flat I_{V_0} + A$ is monotone. Notice that $\mu_\flat^{-1}$ is a time scale. If in addition to (76.10) there exists $\mu_\sharp > \mu_\flat$ so that $\mu_\sharp I_{V_0} + A$ is maximal monotone, then the following result shows that one can extend the Hille–Yosida theorem to the problem $\partial_t u + A(u) = f$, $u(0) = u_0$.

Proposition 76.13 (Well-posedness with weaker monotonicity assumption). *Let $A : V_0 \to L$ satisfy the weaker monotonicity assumption* (76.10)*. Assume that there is a real number $\tau_\flat \in (0, \mu_\flat^{-1})$ s.t. $I_{V_0} + \tau_\flat A$ is surjective. Let $f \in C^1(\overline{J};L)$ and $u_0 \in V_0$. There exists a unique $u \in C^1(\overline{J};L) \cap C^0(\overline{J};V_0)$ solving the problem* (76.6)*. Moreover, introducing the time scale $\rho := (\mu_\flat + \frac{1}{2T})^{-1}$, we have for all $t \in \overline{J}$,*

$$\|u(t)\|_L \leq e^{\frac{t}{\rho}} (tT)^{\frac{1}{2}} \|f\|_{C^0([0,t];L)} + e^{\mu_\flat t} \|u_0\|_L. \tag{76.11}$$

In particular, for $t = T$ we have $\|u(T)\|_L \leq e^{\frac{T}{\rho}} T \|f\|_{C^0(\overline{J};L)} + e^{\mu_\flat T} \|u_0\|_L$.

Proof. Let us set $A_\sharp := \mu_\flat I_{V_0} + A : V_0 \to L$. By assumption, $A_\sharp$ is a monotone operator. Moreover, setting $\tau_\sharp := \frac{\tau_\flat}{1-\tau_\flat\mu_\flat}$ (this is legitimate since $\tau_\flat \mu_\flat < 1$ by assumption), we have

$$I_{V_0} + \tau_\sharp A_\sharp = (1+\tau_\sharp\mu_\flat)\left(I_{V_0} + \frac{\tau_\sharp}{1+\tau_\sharp\mu_\flat} A\right) = \frac{1}{1-\tau_\flat\mu_\flat}(I_{V_0} + \tau_\flat A),$$

where we used that $\tau_\flat = \frac{\tau_\sharp}{1+\tau_\sharp\mu_\flat}$ and $1+\tau_\sharp\mu_\flat = \frac{\tau_\sharp}{\tau_\flat} = \frac{1}{1-\tau_\flat\mu_\flat}$. This shows that the operator $I_{V_0} + \tau_\sharp A_\sharp$ is surjective. In conclusion, $A_\sharp : V_0 \to L$ is a maximal monotone operator. Owing to the Hille–Yosida theorem, there is a unique $v \in C^1(\overline{J};L) \cap C^0(\overline{J};V_0)$ s.t. $v(0) = u_0$ and $\partial_t v(t) + A_\sharp(v(t)) = e^{-\mu_\flat t} f(t)$ for all $t \in J$. Setting $u(t) := e^{\mu_\flat t} v(t)$, we have $u \in C^1(\overline{J};L) \cap C^0(\overline{J};V_0)$, $u(0) = u_0$, and a direct calculation shows that

$$\begin{aligned}\partial_t u(t) &= e^{\mu_\flat t}\partial_t v(t) + \mu_\flat u(t) = e^{\mu_\flat t}(-A_\sharp(v(t)) + e^{-\mu_\flat t} f(t)) + \mu_\flat u(t)\\ &= -A_\sharp(u(t)) + f(t) + \mu_\flat u(t) = -A(u(t)) + f(t).\end{aligned}$$

Hence, u solves (76.6). In addition, the a priori estimate (76.11) follows from $\|u(t)\|_L = e^{\mu_\flat t}\|v(t)\|_L$ and by applying the a priori estimate (76.7) to v. Finally, uniqueness follows from the a priori estimate. □

Remark 76.14 (Time growth). If $\mu_\flat \le \frac{1}{T}$, the a priori estimate (76.11) implies that $\|u(t)\|_L \le e^{\frac{3t}{2T}}(tT)^{\frac{1}{2}}\|f\|_{C^0([0,t];L)} + e^{\frac{t}{T}}\|u_0\|_L$ for all $t \in \overline{J}$, which exhibits essentially the same behavior in time as the a priori estimate (76.7). Instead, if $\mu_\flat \gg \frac{1}{T}$, the factors $e^{\frac{t}{\rho}}$ and $e^{\mu_\flat t}$ in (76.11) can become very large as $t \uparrow T$ (notice that $\frac{1}{\rho} \ge \mu_\flat$ so that $e^{\frac{T}{\rho}} \ge e^{\mu_\flat T}$). □

76.3 Time-dependent Friedrichs' systems

Let us illustrate the above framework with the theory of *Friedrichs' systems* introduced in Chapter 56. We make the assumptions (56.1a)-(56.1b) (boundedness and symmetry), but we do not make the positivity assumption (56.1c). Specifically, let $m \ge 1$ and let $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ be a family of fields in $L^\infty(D;\mathbb{C}^{m\times m})$. We assume that $\mathcal{A}^k = (\mathcal{A}^k)^{\mathsf{H}}$ for all $k \in \{1{:}d\}$ and a.e. in D. Defining $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k \mathcal{A}^k \in L^\infty(D;\mathbb{C}^{m\times m})$ we further assume that $\mathcal{X} \in L^\infty(D;\mathbb{C}^{m\times m})$. Notice that $\mathcal{X}$ is Hermitian. We consider the first-order differential operator

$$A_1(v) := \sum_{k\in\{1:d\}} \mathcal{A}^k \partial_k v. \tag{76.12}$$

As in Remark 56.12, we specify the zero-order operator by means of an operator $K \in \mathcal{L}(L;L)$ with $L := L^2(D;\mathbb{C}^m)$. A simple example is $K(v) := \mathcal{K}v$ with a field $\mathcal{K} \in L^\infty(D;\mathbb{C}^{m\times m})$. Here, K is local, but it is not a requirement. For instance, the Boltzmann equation and the neutron transport equation can be formulated as Friedrichs' systems where the collision operator K is nonlocal.

Let us define the graph space $V := \{v \in L \mid A_1(v) \in L\}$. Proposition 56.4 shows that V is a Hilbert space when equipped with the inner product $(v,w)_L + \ell_D^2\beta^{-2}(A_1(v),A_1(w))_L$. Here, to be dimensionally consistent, we introduced the length scale $\ell_D := \mathrm{diam}(D)$ and the real number $\beta := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(D;\mathbb{C}^{m\times m})}$. Following (56.26), we define the self-adjoint boundary operator $N \in \mathcal{L}(V;V')$ s.t.

$$\langle N(v), w\rangle_{V',V} := (\mathcal{X}v, w)_L + (A_1(v), w)_L + (v, A_1(w))_L. \tag{76.13}$$

It has been shown in Chapter 56 that the boundary conditions for Friedrichs' systems can be formulated by postulating the existence of a monotone operator $M \in \mathcal{L}(V;V')$ such that $\ker(N-M) + \ker(N+M) = V$ (see §56.3.2). Homogeneous boundary conditions can be enforced by considering $V_0 := \ker(M-N)$. Notice that V_0 is a closed subspace of V. Moreover, since $C_0^\infty(D;\mathbb{C}^m)$ is dense in L (see Theorem 1.38), the inclusions $C_0^\infty(D;\mathbb{C}^m) \subset V_0 \subset V \subset L$ imply that V_0 and V are dense in L.

Let us define the operator $A : V_0 \to L$ such that

$$A(v) := K(v) + A_1(v), \tag{76.14}$$

and let us consider the time evolution problem (76.6), i.e., we seek $u \in C^1(\overline{J};L) \cap C^0(\overline{J};V_0)$ s.t. $\partial_t u(t) + A(u(t)) = f(t)$ for all $t \in \overline{J}$, and $u(0) = u_0$, with $f \in C^1(\overline{J};L)$ and $u_0 \in V_0$. To be dimensionally consistent (see Remark 76.2), the operators A_1 and K have both the dimension of the reciprocal of a time. This implies that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ and the real number β have the

dimension of a velocity. The following real number will play an important role in the forthcoming stability and error analysis:

$$\Lambda_\flat := \inf_{v\in L} \frac{\Re\big((K(v),v)_L\big) - \frac{1}{2}(\mathcal{X}v,v)_L}{\|v\|_L^2}. \tag{76.15}$$

Notice that $|\Lambda_\flat|^{-1}$ is a time scale, and $\Lambda_\flat$ is indeed real since $\mathcal{X}$ is Hermitian. Setting $\Lambda_\flat^- := \max(0,-\Lambda_\flat) \ge 0$, we introduce the time scale

$$\rho := (2\Lambda_\flat^- + \beta\ell_D^{-1})^{-1}. \tag{76.16}$$

Proposition 76.15 (Well-posedness for Friedrichs' systems). *Let A, V_0, and L be defined as above.* (i) *For all $f \in C^1(\overline{J};L)$ and all $u_0 \in V_0$, there exists a unique solution u to* (76.6). (ii) *The solution satisfies the a priori estimate* (76.7) *if $\Lambda_\flat = 0$, the a priori estimate* (76.9) *with $\mu_\sharp := \Lambda_\flat$ if $\Lambda_\flat > 0$, and the a priori estimate* (76.11) *with $\mu_\flat := -\Lambda_\flat$ if $\Lambda_\flat < 0$.*

Proof. The definition (76.13) of N implies that $\Re\big((A_1(v),v)_L\big) = -\frac{1}{2}(\mathcal{X}v,v)_L + \frac{1}{2}\langle N(v),v\rangle_{V',V}$ for all $v\in V_0$. Since $V_0 = \ker(N-M)$ and M is a monotone operator, we infer that

$$\begin{aligned}\Re\big((A(v),v)_L\big) &\ge \Lambda_\flat\|v\|_L^2 + \frac{1}{2}\langle N(v),v\rangle_{V',V}\\ &= \Lambda_\flat\|v\|_L^2 + \frac{1}{2}\langle M(v),v\rangle_{V',V} \ge \Lambda_\flat\|v\|_L^2.\end{aligned}$$

If $\Lambda_\flat \ge 0$, the above lower bound shows that the operator A is monotone, whereas if $\Lambda_\flat < 0$, the operator A satisfies the weak monotonicity property (76.10). Moreover, the theory of Friedrichs' systems shows that for any real number $\rho > \max(0,-\Lambda_\flat) > 0$, the operator $B := \rho I_{V_0} + A : V_0 \to L$ is an isomorphism (see Theorem 56.9). This implies that the operator $I_{V_0} + \tau_0 A : V_0 \to L$ is surjective for any $\tau_0 \in (0,\rho^{-1})$. We conclude by applying the Hille–Yosida theorem if $\Lambda_\flat \ge 0$ (see also see Exercise 76.3) or its variant stated in Proposition 76.13 if $\Lambda_\flat < 0$. □

76.4 Space semi-discretization

In this section, we present the space discretization of the model problem (76.6) with H^1-conforming finite elements. We enforce the boundary condition by means of the boundary penalty method introduced in §57.4 and we use the fluctuation-based stabilization techniques presented in Chapters 58-59. We consider the setting of the time-dependent Friedrichs' systems from §76.3. The operator A is defined in (76.14) and we assume that all the assumptions stated in §76.3 for the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ and the operator $\mathcal{K}$ hold true. The fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are also assumed to be Lipschitz in D.

76.4.1 Discrete setting

As in §58.1, we consider a shape-regular mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly, and for all $h\in\mathcal{H}$ we consider an H^1-conforming finite-dimensional subspace $V_h \subset V$ built by using a finite element of degree $k\ge 1$. To simplify the argumentation, $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is assumed to be quasi-uniform, and the typical meshsize of $\mathcal{T}_h$ is denoted by h. From now on, we assume that there is $s > \frac{1}{2}$ so that the solution to (76.6) is s.t. $u\in C^0(\overline{J},H^s(D;\mathbb{C}^m))$, and we set $V_\sharp := H^s(D;\mathbb{C}^m)+V_h$. Let $\mathcal{P}_{V_h} : L \to V_h$ be the L-orthogonal projection onto V_h, i.e., for all $z\in L$, $\mathcal{P}_{V_h}(z)$ is the unique element in V_h s.t. $(z-\mathcal{P}_{V_h}(z),w_h)_L = 0$ for all $w_h\in V_h$.

We enforce the boundary condition by means of the boundary penalty method introduced in §57.4. Setting $L(\partial D) := L^2(\partial D;\mathbb{C}^m)$, we assume that there are boundary fields

$$\mathcal{M},\mathcal{N} \in L^\infty(\partial D;\mathbb{C}^{m\times m})$$

s.t. for all $v, w \in V_\sharp$,

$$\langle M(v), w\rangle_{V',V} = (\mathcal{M}v, w)_{L(\partial D)}, \quad \langle N(v), w\rangle_{V',V} = (\mathcal{N}v, w)_{L(\partial D)}. \tag{76.17}$$

We also introduce a boundary penalty field $\mathcal{S}^\partial \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ taking values over the set of the $m\times m$ Hermitian positive semidefinite matrices, and we set

$$\mathcal{M}^{\text{BP}} := \mathcal{M} + \mathcal{S}^\partial, \qquad |v|_{\mathcal{M}^{\text{BP}}} := \frac{1}{2}(\mathcal{M}^{\text{BP}}v, v)^{\frac12}_{L(\partial D)}, \quad \forall v \in V_\sharp. \tag{76.18}$$

Setting $\beta := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(D;\mathbb{C}^{m\times m})}$, we assume that $\mathcal{S}^\partial$ is defined in such a way that there is c s.t.

$$\ker(\mathcal{M}-\mathcal{N}) \subset \ker(\mathcal{M}^{\text{BP}}-\mathcal{N}), \tag{76.19a}$$

$$|w_h|_{\mathcal{M}^{\text{BP}}} \le c\,\big(\tfrac{\beta}{h}\big)^{\frac12}\|w_h\|_L, \tag{76.19b}$$

$$|((\mathcal{M}^{\text{BP}}+\mathcal{N})v, w)_L| \le c\,\beta^{\frac12}\|v\|_{L(\partial D)}|w|_{\mathcal{M}^{\text{BP}}}, \tag{76.19c}$$

for all $v, w \in V_\sharp$, all $w_h \in V_h$, and all $h \in \mathcal{H}$. We have shown in §57.4.2 how to construct boundary penalty fields $\mathcal{S}^\partial$ satisfying (57.33). Then (76.19a) and (76.19c) are satisfied since they are restatements of (57.33a) and (57.33d), respectively. Moreover, (76.19b) is also satisfied as a consequence of (57.33b) and a discrete trace inequality.

We now introduce a Hermitian semidefinite stabilization sesquilinear form s_h on $V_h\times V_h$. We can use the continuous interior penalty (CIP) stabilization described in §58.3 or the two-scale stabilization techniques described in Chapter 59, i.e., the local projection stabilization (LPS) or the subgrid viscosity (SGV) methods. Since we do not want to be specific about the type of fluctuation-based stabilization we use, we introduce generic properties that s_h should satisfy. Setting $|v_h|_{\mathcal{S}} := s_h(v_h, v_h)^{\frac12}$, we assume that the following simplified assumptions stated in Remark 58.1 hold true: There is c s.t.

$$|w_h|_{\mathcal{S}} \le c\,\big(\tfrac{\beta}{h}\big)^{\frac12}\|w_h\|_L, \tag{76.20a}$$

$$|s_h(\mathcal{P}_{V_h}(v), w_h)| \le c\,\beta^{\frac12}h^{k+\frac12}|v|_{H^{k+1}(D;\mathbb{C}^m)}|w_h|_{\mathcal{S}}, \tag{76.20b}$$

$$|(v-\mathcal{P}_{V_h}(v), A_1(w_h))_L| \le c\,\beta^{\frac12}h^{k+\frac12}|v|_{H^{k+1}(D;\mathbb{C}^m)} \times \Big(|w_h|_{\mathcal{S}} + \big(\tfrac{\beta}{\ell_D}\big)^{\frac12}\|w_h\|_L\Big), \tag{76.20c}$$

for all $v \in H^{k+1}(D;\mathbb{C}^m)$, all $w_h \in V_h$, and all $h \in \mathcal{H}$. In what follows, we assume that s_h is defined in (58.24) or (58.25) for CIP, in (59.13) or (59.14) for LPS, or in (59.19) or (59.20) for SGV. Then the assumptions (76.20) are met, as shown in Remark 58.1, in Remark 58.12, and in Remark 59.12.

Remark 76.16 (Variants). Stabilization by discontinuous Galerkin methods can be considered as well but it is not discussed here for brevity. Residual-based stabilization techniques could be also used, but they introduce additional technicalities because the residual depends on the time derivative of the solution. The reader is referred to Johnson et al. [201], Burman [59] for results in this direction. □

76.4.2 Discrete problem and well-posedness

We define the sesquilinear form a_h on $V_h \times V_h$ such that

$$a_h(v_h, w_h) := (A(v_h), w_h)_L + \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} - \mathcal{N})v_h, w_h)_{L(\partial D)} + s_h(v_h, w_h). \tag{76.21}$$

Proceeding as in the proof of Proposition 76.15, we obtain the following important lower bound on the sesquilinear form a_h: For all $v_h \in V_h$,

$$\Re\big(a_h(v_h, v_h)\big) \geq \Lambda_\flat \|v_h\|_L^2 + |v_h|_{\mathcal{MS}}^2, \quad |v_h|_{\mathcal{MS}}^2 := |v_h|_{\mathcal{M}^{\mathrm{BP}}}^2 + |v_h|_{\mathcal{S}}^2. \tag{76.22}$$

The space semi-discrete problem is as follows:

$$\begin{cases} \text{Find } u_h \in C^1(\overline{J}; V_h) \text{ s.t. } u_h(0) = \mathcal{P}_{V_h}(u_0) \text{ and} \\ (\partial_t u_h(t), w_h)_L + a_h(u_h(t), w_h) = (f(t), w_h)_L, \ \forall t \in \overline{J}, \ \forall w_h \in V_h. \end{cases} \tag{76.23}$$

Setting $f_h(t) := \mathcal{P}_{V_h}(f(t))$ for all $t \in \overline{J}$ and defining $A_h : V_h \to V_h$ s.t. $(A_h(v_h), w_h)_L := a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$, (76.23) becomes

$$\begin{cases} \text{Find } u_h \in C^1(\overline{J}; V_h) \text{ s.t. } u_h(0) = \mathcal{P}_{V_h}(u_0) \text{ and} \\ \partial_t u_h(t) + A_h(u_h(t)) = f_h(t), \qquad \forall t \in \overline{J}. \end{cases} \tag{76.24}$$

Proposition 76.17 (Well-posedness, a priori estimate). (i) *The semi-discrete problem* (76.23) *is well-posed.* (ii) u_h *satisfies the a priori estimate* (76.7) *if* $\Lambda_\flat = 0$*, the a priori estimate* (76.9) *with* $\mu_\sharp := \Lambda_\flat$ *if* $\Lambda_\flat > 0$*, and the a priori estimate* (76.11) *with* $\mu_\flat := -\Lambda_\flat$ *if* $\Lambda_\flat < 0$*.*

Proof. Since V_h is finite-dimensional, the well-posedness of (76.23) follows from the Cauchy–Lipschitz theorem. The a priori estimates follow by observing that $\|u_h(0)\|_L \leq \|u_0\|_L$ since $\|\mathcal{P}_{V_h}(u_0)\|_L \leq \|u_0\|_L$, and by proceeding as in the proof of the Hille–Yosida theorem if $\Lambda_\flat = 0$, as in Exercise 76.3 if $\Lambda_\flat > 0$, and as in the proof of Proposition 76.13 if $\Lambda_\flat < 0$. □

76.4.3 Error analysis

The starting point of the error analysis is the identity $(\partial_t u_h(t), w_h)_L + a_h(u_h(t), w_h) = (f(t), w_h)_L = (\partial_t u(t), w_h)_L + (A(u(t)), w_h)_L$ for all $w_h \in V_h$. Let $v_h \in C^1(\overline{J}; V_h)$ and set $e_h := u_h - v_h$. We then obtain the following error equation: For all $t \in \overline{J}$,

$$(\partial_t e_h, w_h)_L + a_h(e_h, w_h) = \big((A(u), w_h)_L - a_h(v_h, w_h)\big) + (\partial_t(u - v_h), w_h)_L.$$

There are essentially two possibilities to proceed from here depending on the way one wants to handle the two terms on the right-hand side.

In the first approach, one chooses v_h so as to eliminate the consistency error induced by the approximation of the differential operator A. This is the most natural approach in the context of the Hille–Yosida theorem. Recalling the time scale ρ defined in (76.16), we are led to define the approximation operator $\Pi_h^{\mathrm{A}} : V \to V_h$ s.t.

$$\rho^{-1}(\Pi_h^{\mathrm{A}}(v), w_h)_L + a_h(\Pi_h^{\mathrm{A}}(v), w_h) = \rho^{-1}(v, w_h)_L + (A(v), w_h)_L, \tag{76.25}$$

for all $v \in V$ and all $w_h \in V_h$. This problem is well-posed since the sesquilinear form $\tilde{a}_h(v_h, w_h) := \rho^{-1}(v_h, w_h)_L + a_h(v_h, w_h)$ satisfies an inf-sup condition on $V_h \times V_h$; see Lemma 58.2. Notice that $\tilde{a}_h$ is L-coercive on V_h with the constant $\mu_0 := \rho^{-1} + \Lambda_\flat \geq \beta \ell_D^{-1} + \Lambda_\flat^- \geq \frac{1}{2}\rho^{-1}$. The approximation

properties of Π_h^{A} are stated in Theorem 58.11 for CIP and in Theorem 59.11 for LPS and SGV. Setting $e_h := u_h - \Pi_h^{\mathrm{A}}(u)$, $\eta := \Pi_h^{\mathrm{A}}(u) - u$, the error equation becomes

$$(\partial_t e_h, w_h)_L + a_h(e_h, w_h) = \rho^{-1}(\eta, w_h)_L - (\partial_t \eta, w_h)_L. \tag{76.26}$$

The second choice is to use the all-purpose L-orthogonal projection operator $\mathcal{P}_{V_h}$ to eliminate the consistency error on the time derivative. Recall that $\mathcal{P}_{V_h}$ has optimal approximation properties in L^2 and H^1 since we are using quasi-uniform mesh sequences; see Propositions 22.19 and 22.21. Setting $v_h := \mathcal{P}_{V_h}(u)$, $e_h := u_h - \mathcal{P}_{V_h}(u)$, $\eta := \mathcal{P}_{V_h}(u) - u$, we observe that $(\partial_t \eta, w_h)_L = 0$ since $\partial_t(\mathcal{P}_{V_h}(u)) = \mathcal{P}_{V_h}(\partial_t u)$. The error then equation becomes

$$\begin{aligned}(\partial_t e_h, w_h)_L + a_h(e_h, w_h) = (\eta, A_1(w_h))_L - (K(\eta) - \mathcal{X}\eta, w_h)_L \\ - s_h(\mathcal{P}_{V_h}(u), w_h) - \frac{1}{2}((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta, w_h)_{L(\partial D)},\end{aligned} \tag{76.27}$$

where we used that $A(\eta) = K(\eta) + A_1(\eta)$ (see (76.14)), the integration by parts formula (76.13), and (76.19a) (which implies that $(\mathcal{M}^{\mathrm{BP}} - \mathcal{N})u = 0$).

We now derive an $L^\infty(\overline{J}; L)$-error estimate using the above two approaches. To simplify the tracking of model-dependent constants, we set $\check{c} := \rho \max(\|K\|_{\mathcal{L}(L;L)}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, L_{\mathcal{A}})$, where $L_{\mathcal{A}}$ is the Lipschitz constant of the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$, and we hide the constant $\check{c}$ in the generic constants used in the error analysis. (Notice that $\|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})} \le dL_{\mathcal{A}}$.) We also use the nondimensional function $\xi(t) := \min(\frac{2t}{\rho}, 1)$ for all $t \in \overline{J}$.

Theorem 76.18 ($L^\infty(\overline{J}; L)$-estimate using Π_h^{A}). *Let u solve (76.6) and let u_h solve (76.23). Assume that $u \in C^1(\overline{J}; H^{k+1}(D;\mathbb{C}^m))$. There is c s.t. for all $h \in \mathcal{H}$ and all $t \in \overline{J}$, with $J_t := (0,t)$,*

$$\|u - u_h\|_{L^\infty(\overline{J}_t;L)} \le c\, e^{\frac{t}{\rho}} \max(\rho\beta, h)^{\frac{1}{2}} h^{k+\frac{1}{2}} c_1(t;u), \tag{76.28}$$

with $c_1(t;u) := |u|_{C^0([0,t];H^{k+1}(D;\mathbb{C}^m))} + \xi(t)^{\frac{1}{2}}\rho|\partial_t u|_{C^0([0,t];H^{k+1}(D;\mathbb{C}^m))}$.

Proof. Using $w_h := e_h$ in (76.26), taking the real part, and using the lower bound (76.22) on the sesquilinear form a_h yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_L^2 + \Lambda_\flat \|e_h\|_L^2 + |e_h|_{\mathcal{MS}}^2 \le \rho^{-1}\Phi_\eta^2 + \frac{1}{2}\rho^{-1}\|e_h\|_L^2,$$

with $\Phi_\eta^2 := \|\eta\|_L^2 + \rho^2\|\partial_t\eta\|_L^2$ and where we used the Cauchy–Schwarz inequality and Young's inequality on the right-hand side. Since $-\Lambda_\flat \le \Lambda_\flat^-$, we have $\frac{1}{2}\rho^{-1} - \Lambda_\flat \le \rho^{-1}$. Dropping the nonnegative term $|e_h|_{\mathcal{MS}}^2$ from the left-hand side, we infer that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_L^2 \le 2\rho^{-1}\|e_h\|_L^2 + 2\rho^{-1}\Phi_\eta^2.$$

Invoking a simplified form of Gronwall's lemma (see Exercise 73.2), this bound implies that for all $t \in \overline{J}$,

$$\|e_h\|_{L^\infty(\overline{J}_t;L)}^2 \le e^{\frac{2t}{\rho}}\Big(\|e_h(0)\|_L^2 + \xi(t)\|\Phi_\eta\|_{L^\infty(\overline{J}_t)}^2\Big).$$

Taking the square root, invoking the triangle inequality, and recalling the definition of Φ_η yields

$$\|u - u_h\|_{L^\infty(\overline{J}_t;L)} \le e^{\frac{t}{\rho}}\big(\|e_h(0)\|_L + 2\|\eta\|_{L^\infty(\overline{J}_t;L)} + \xi(t)^{\frac{1}{2}}\rho\|\partial_t\eta\|_{L^\infty(\overline{J}_t;L)}\big).$$

Applying Theorem 58.11 (for CIP) or Theorem 59.11 (for LPS or SGV) with $\mu_0 := \frac{1}{2\rho}$ gives $\|v - \Pi_h^{\mathrm{A}}(v)\|_L \le c\max(\rho\beta, h)^{\frac{1}{2}} h^{k+\frac{1}{2}}|v|_{H^{k+1}}$ for all $v \in H^{k+1}(D;\mathbb{C}^m)$. This estimate allows us

to bound $\|\eta\|_{L^\infty(\overline{J}_t;L)}$. Moreover, since the differential operator A is time-independent, we have $\partial_t(\Pi_h^{\mathrm{A}}(u)) = \Pi_h^{\mathrm{A}}(\partial_t u)$, and we can bound $\|\partial_t\eta\|_{L^\infty(\overline{J}_t;L)}$ similarly. Altogether this yields

$$2\|\eta\|_{L^\infty(\overline{J}_t;L)} + \xi(t)^{\frac12}\rho\|\partial_t\eta\|_{L^\infty(\overline{J}_t)} \le c\,\max(\rho\beta,h)^{\frac12}h^{k+\frac12}c_1(t;u).$$

Finally, we have $\|e_h(0)\|_L \le \|(I_L - \mathcal{P}_{V_h})(u(0))\|_L + \|\eta(0)\|_L \le 2\|\eta(0)\|_L$, and $\|\eta(0)\|_L$ has been bounded above. This completes the proof. □

Theorem 76.19 ($L^\infty(\overline{J};L)$-estimate using $\mathcal{P}_{V_h}$). *Let u solve* (76.6) *and let u_h solve* (76.23). *Assume that $u \in C^0(\overline{J};H^{k+1}(D;\mathbb{C}^m))$. There is c s.t. for all $h \in \mathcal{H}$ and all $t \in \overline{J}$, with $J_t := (0,t)$,*

$$\|u - u_h\|_{L^\infty(\overline{J}_t;L)} \le c\left(h^{\frac12} + e^{\frac{t}{\rho}}\xi(t)^{\frac12}\max(\rho\beta,h)^{\frac12}\right)h^{k+\frac12}c_2(t;u), \tag{76.29}$$

with $c_2(t;u) := |u|_{C^0([0,t];H^{k+1}(D;\mathbb{C}^m))}$.

Proof. Using $w_h := e_h$ in (76.26), taking the real part, and using the lower bound (76.22) on the sesquilinear form a_h yields

$$\frac12\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_L^2 + \Lambda_\flat\|e_h\|_L^2 + |e_h|_{\mathcal{MS}}^2 \le |\Phi_\eta(e_h)|,$$

with the antilinear form $\Phi_\eta \in V_h'$ s.t. $\Phi_\eta(w_h) := (\eta, A_1(w_h))_L - (K(\eta) - \mathcal{X}\eta, w_h)_L - s_h(\mathcal{P}_{V_h}(u), w_h) - \frac12((\mathcal{M}^{\mathrm{BP}} + \mathcal{N})\eta, w_h)_{L(\partial D)}$ for all $w_h \in V_h$. Let us equip the space V_h with the norm $\|v_h\|_{V_\flat}^2 := \rho^{-1}\|v_h\|_L^2 + |v_h|_{\mathcal{MS}}^2$, and let us set $\|\Phi_\eta\|_{V_{h\flat}'} := \sup_{w_h\in V_h}\frac{|\Phi_\eta(w_h)|}{\|w_h\|_{V_\flat}}$. Since

$$|\Phi_\eta(e_h)| \le \frac12\|\Phi_\eta\|_{V_{h\flat}'}^2 + \frac12\|e_h\|_{V_\flat}^2 = \frac12\|\Phi_\eta\|_{V_{h\flat}'}^2 + \frac{1}{2\rho}\|e_h\|_L^2 + \frac12|e_h|_{\mathcal{MS}}^2,$$

proceeding as in the previous proof leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_L^2 \le 2\rho^{-1}\|e_h\|_L^2 + \|\Phi_\eta\|_{V_{h\flat}'}^2,$$

so that for all $t \in \overline{J}$,

$$\|e_h\|_{L^\infty(\overline{J}_t;L)}^2 \le e^{\frac{2t}{\rho}}\xi(t)\tfrac12\rho\|\Phi_\eta\|_{L^\infty(\overline{J}_t;V_{h\flat}')}^2,$$

where we used that $e_h(0) = 0$. Taking the square root and invoking the triangle inequality gives

$$\|u - u_h\|_{L^\infty(\overline{J}_t;L)} \le \|\eta\|_{L^\infty(\overline{J}_t;L)} + e^{\frac{t}{\rho}}\tfrac{1}{\sqrt{2}}\xi(t)\rho^{\frac12}\|\Phi_\eta\|_{L^\infty(\overline{J}_t;V_{h\flat}')}.$$

The approximation properties of $\mathcal{P}_{V_h}$ in L^2 imply that $\|\eta\|_{L^\infty(\overline{J}_t;L)} \le ch^{k+1}c_2(u)$. To bound $\|\Phi_\eta\|_{V_{h\flat}'}$ for all $t \in \overline{J}$, we invoke (76.20c) and $\frac{\beta}{\ell_D} \le \rho^{-1}$ for the first term, the Cauchy–Schwarz inequality for the second term, (76.20b) for the third term, and (76.19c) for the fourth term. Using the approximation properties of $\mathcal{P}_{V_h}$ in L^2 and H^1 (see Propositions 22.19 and 22.21), we infer that

$$\rho^{\frac12}\|\Phi_\eta\|_{V_{h\flat}'} \le c\,\max(\rho\beta,h)^{\frac12}h^{k+\frac12}|u|_{H^{k+1}(D;\mathbb{C}^m)}.$$

Putting everything together yields the assertion. □

Remark 76.20 (Comparison). The estimates from Theorem 76.18 and Theorem 76.19 lead to the same decay rates in h. Using the operator Π_h^{A} is more natural in the setting of the Hille–Yosida theorem. This is reflected by the fact that the proof of Theorem 76.18 is simpler since

it does not invoke the structural assumptions related to the boundary penalty method and the fluctuation-based stabilization. Indeed, the proof just uses the approximation properties of Π_h^{A} which have been established in Theorems 58.11 and 59.11 using these assumptions (notice that these approximation properties only require shape-regular mesh sequences). The proof of Theorem 76.19 goes through these arguments again when bounding $\|\Phi_\eta\|_{V_{h\flat}'}$. Theorem 76.18, however, requires a stronger regularity assumption than Theorem 76.19 since it assumes that $u \in C^1(\overline{J}; H^{k+1}(D;\mathbb{C}^m))$ instead of $u \in C^0(\overline{J}; H^{k+1}(D;\mathbb{C}^m))$. Moreover, Theorem 76.18 uses that the differential operator is time-independent. Finally, Theorem 76.19 relies on the approximation properties in H^1 of $\mathcal{P}_{V_h}$, but these properties are more delicate to establish beyond the setting of quasi-uniform mesh sequences (see Remark 22.23). □

Exercises

Exercise 76.1 (Maximality). Let $V \hookrightarrow L$ be two real Hilbert spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_L$. Let $R \in \mathcal{L}(V;L)$. Assume that R is a monotone operator, i.e., $\Re((R(v),v)_L) \geq 0$ for all $v \in V$. (i) Show that if R is maximal monotone (i.e., there is $\tau_0 > 0$ s.t. $I_V + \tau_0 R$ is surjective), then there are real numbers $c_1 > 0$ and $c_2 > 0$ s.t. $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \geq c_1\|v\|_V - c_2\|v\|_L$ for all $v \in V$. (*Hint*: show that $I_V + \tau_0 R$ is injective with closed image.) (ii) Show that if there are real numbers $c_1 > 0$ and $c_2 > 0$ s.t. $\sup_{w\in L} \frac{|(R(v),w)_L|}{\|w\|_L} \geq c_1\|v\|_V - c_2\|v\|_L$ for all $v \in V$, and $c_2 I_L + R^* : L' \equiv L \to V'$ is injective, then R is maximal monotone. (*Hint*: consider $S(v) := \sup_{w\in L} \frac{|(R(v)+c_2 v,w)_L|}{\|w\|_L}$ for all $v \in V$.) (iii) Assume that $I_V + \tau_0 R$ is surjective. Show that the norms $\|v\|_L + \tau_0\|R(v)\|_L$ and $\|v\|_V$ are equivalent.

Exercise 76.2 (Lemma 76.8). Revisit the proof of Lemma 76.8 by using Young's inequality in the form $a(s)\phi(s)^{\frac{1}{2}} \leq \frac{\theta a(s)^2}{4} + \frac{\phi(s)}{\theta}$, where θ is any time scale, and show that the choice $\theta = T$ leads to the sharpest estimate at the final time $t = T$. (*Hint*: minimize the function $\theta \mapsto \theta e^{\frac{T}{\theta}}$ at fixed T.)

Exercise 76.3 (Growth and decay in time). Assume that the linear operator $-\mu_\flat I_L + A \in \mathcal{L}(V_0;L)$ is maximal monotone where $\mu_\flat \in \mathbb{R}$, $\mu_\flat \neq 0$, but there is no constraint on the sign of $\mu_\flat$. Let $f \in C^0(\mathbb{R}_+;L)$ $\mathbb{R}_+ := [0,\infty)$. (i) Explain why there exists a unique $u \in C^1(\mathbb{R}_+;V_0)\cap C^0(\mathbb{R}_+;V_0)$ solving the problem $\partial_t u + A(u) = f$, $u(0) = u_0$. (ii) Assume now that $\mu_\flat > 0$. Show that the solution to this problem satisfies the following estimate for all $t \geq 0$:

$$\|u(t)\|_L^2 \leq e^{-\mu_\flat t}\|u_0\|_L^2 + \frac{1}{\mu_\flat}\int_0^t e^{-\mu_\flat(t-s)}\|f(s)\|_L^2 \,\mathrm{d}s.$$

(iii) Assume that $\mu_\flat > 0$ and $f \in C^0(\mathbb{R}_+;L) \cap L^\infty((0,\infty);L)$. Show that $\limsup_{t\to\infty} \|u(t)\|_L \leq \mu_\flat^{-1}\|f\|_{L^\infty((0,\infty);L)}$.

Exercise 76.4 (Wave equation). Consider the wave equation $\partial_{tt}p - \Delta p = g$ in $D\times J$ with the initial conditions $p(0) = p_0$ and $\partial_t p(0) = v_0$ in D and homogeneous Dirichlet conditions on p at the boundary. Assume that $g \in L^2(D)$, $p_0, v_0 \in H_0^1(D)$, and $\Delta p_0 \in L^2(D)$. Show that this problem fits the setting of the time-dependent Friedrichs' systems from §76.3. (*Hint*: introduce $v := \partial_t p$ and $\boldsymbol{q} := -\nabla p$.)

Chapter 77

Implicit time discretization

In this chapter, we continue the study of the time-dependent Friedrichs' systems introduced in Chapter 76. In the previous chapter, we established the well-posedness of the continuous model problem (76.6) and we discretized the problem in space using H^1-conforming finite elements, a boundary penalty technique, and fluctuation-based stabilization. In this chapter, we now discretize the problem in time and focus on the implicit Euler scheme. The explicit Euler scheme and explicit Runge–Kutta schemes are investigated in Chapter 78.

77.1 Model problem and space discretization

In this section, we briefly review the model problem under consideration and recall the setting for the space discretization introduced in §76.4.

77.1.1 Model problem

Let D be a Lipschitz domain in $\mathbb{R}^d$ and let $J := (0,T)$ be the time interval with $T > 0$. Let $m \geq 1$ and let $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ be a family of fields in $L^\infty(D;\mathbb{C}^{m\times m})$ such that $\mathcal{X} := \sum_{k\in\{1:d\}} \partial_k \mathcal{A}^k \in L^\infty(D;\mathbb{C}^{m\times m})$, and $\mathcal{A}^k = (\mathcal{A}^k)^{\mathsf{H}}$ for all $k \in \{1{:}d\}$ and a.e. in D. We consider the first-order differential operator $A_1(v) := \sum_{k\in\{1:d\}} \mathcal{A}^k \partial_k v$. Recall that the graph space $V := \{v \in L \mid A_1(v) \in L\}$ is a Hilbert space when equipped with the inner product $(v,w)_L + \ell_D^2 \beta^{-2} (A_1(v), A_1(w))_L$ with the length scale $\ell_D := \operatorname{diam}(D)$ and the velocity scale $\beta := \max_{k\in\{1:d\}} \|\mathcal{A}^k\|_{L^\infty(D;\mathbb{C}^{m\times m})}$. Homogeneous boundary conditions are enforced by considering the subspace $V_0 := \ker(M-N) \subset V$, where $N \in \mathcal{L}(V;V')$ is the self-adjoint boundary operator defined in (76.13) and $M \in \mathcal{L}(V;V')$ is the monotone operator s.t. $\ker(N-M) + \ker(N+M) = V$. Let $K \in \mathcal{L}(L;L)$ and define the Friedrichs' operator $A : V_0 \to L$ such that $A(v) := K(v) + A_1(v)$. Let $f \in C^1(\overline{J};L)$ and $u_0 \in V_0$. The time evolution problem is the following:

$$\begin{cases} \text{Find } u \in C^1(\overline{J};L) \cap C^0(\overline{J};V_0) \text{ s.t. } u(0) = u_0 \text{ and} \\ \partial_t u(t) + A(u(t)) = f(t), \qquad \forall t \in \overline{J}. \end{cases} \tag{77.1}$$

From now on, we assume that $u \in C^0(\overline{J}, H^s(D;\mathbb{C}^m))$ with $s > \frac{1}{2}$.

We introduce the real number

$$\Lambda_{\flat} := \inf_{v\in L} \frac{\Re\big((K(v),v)_L\big) - \frac{1}{2}(\mathcal{X}v,v)_L}{\|v\|_L^2}. \tag{77.2}$$

Notice that $|\Lambda_\flat|^{-1}$ is a time scale. No assumption is made on the sign of $\Lambda_\flat$. As in the previous chapter (see (76.16)), we introduce the time scale

$$\rho := (2\Lambda_\flat^- + \beta\ell_D^{-1})^{-1}, \tag{77.3}$$

where $\Lambda_\flat^- := \max(0, -\Lambda_\flat) \geq 0$. If $\Lambda_\flat \geq 0$, then $\rho := \ell_D\beta^{-1}$. If instead $\Lambda_\flat < 0$, then $\rho < \ell_D\beta^{-1}$. The time scale ρ is invoked repeatedly in this chapter. In particular, we use that $\rho^{-1} + \Lambda_\flat \geq \frac{1}{2}\rho^{-1}$ and $2\rho\Lambda_\flat^- \leq 1$.

77.1.2 Setting for the space discretization

As in §76.4.1, we consider a shape-regular mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly. To simplify the argumentation, $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is assumed to be quasi-uniform, and the typical meshsize of $\mathcal{T}_h$ is denoted by h. For all $h \in \mathcal{H}$, we consider an H^1-conforming finite-dimensional subspace $V_h \subset V$ built by using a finite element of degree $k \geq 1$. We set $V_\sharp := H^s(D;\mathbb{C}^m) + V_h$.

To enforce the boundary condition, we assume that there are boundary fields

$$\mathcal{M}, \mathcal{N} \in L^\infty(\partial D; \mathbb{C}^{m\times m})$$

s.t. $\langle M(v), w\rangle_{V',V} = (\mathcal{M}v, w)_{L(\partial D)}$ and $\langle N(v), w\rangle_{V',V} = (\mathcal{N}v, w)_{L(\partial D)}$ for all $v, w \in V_\sharp$, and $L(\partial D) := L^2(\partial D;\mathbb{C}^m)$. The boundary penalty method is based on a field $\mathcal{S}^\partial \in L^\infty(\partial D;\mathbb{C}^{m\times m})$ taking values over the set of the $m\times m$ Hermitian positive semidefinite matrices. Setting $\mathcal{M}^{\text{BP}} := \mathcal{M} + \mathcal{S}^\partial$ and defining the seminorm $|v|_{\mathcal{M}^{\text{BP}}} := \frac{1}{2}(\mathcal{M}^{\text{BP}}v, v)^{\frac{1}{2}}_{L(\partial D)}$ on $V_\sharp$, we assume that $\mathcal{S}^\partial$ satisfies (76.19). The fluctuation-based stabilization is based on a Hermitian semidefinite bilinear form s_h defined on $V_\sharp\times V_\sharp$. Letting $|v|_{\mathcal{S}} := s_h(v,v)^{\frac{1}{2}}$ for all $v \in V_\sharp$, we assume that s_h satisfies (76.20). We define the sesquilinear form a_h on $V_h\times V_h$ such that

$$a_h(v_h, w_h) := (A(v_h), w_h)_L + \frac{1}{2}((\mathcal{M}^{\text{BP}} - \mathcal{N})v_h, w_h)_{L(\partial D)} + s_h(v_h, w_h). \tag{77.4}$$

An important property of a_h is that for all $v_h \in V_h$,

$$\Re\big(a_h(v_h, v_h)\big) \geq \Lambda_\flat\|v_h\|_L^2 + |v_h|_{\mathcal{MS}}^2, \tag{77.5}$$

with $|v|_{\mathcal{MS}}^2 := |v|_{\mathcal{M}^{\text{BP}}}^2 + |v|_{\mathcal{S}}^2$. As we did in the space semi-discrete case, we assume that the fields $\{\mathcal{A}^k\}_{k\in\{1:d\}}$ are Lipschitz with constant $L_{\mathcal{A}}$, we set

$$\check{c} := \rho\max(\|K\|_{\mathcal{L}(L;L)}, \|\mathcal{X}\|_{L^\infty(D;\mathbb{C}^{m\times m})}, L_{\mathcal{A}}), \tag{77.6}$$

and we hide the constant $\check{c}$ in the generic constants used in the error analysis.

The space semi-discrete problem is as follows:

$$\begin{cases} \text{Find } u_h \in C^1(\overline{J}; V_h) \text{ s.t. } u_h(0) = \mathcal{P}_{V_h}(u_0) \text{ and} \\ (\partial_t u_h(t), w_h)_L + a_h(u_h(t), w_h) = (f(t), w_h)_L,\ \forall t \in \overline{J},\ \forall w_h \in V_h. \end{cases} \tag{77.7}$$

Setting $f_h(t) := \mathcal{P}_{V_h}(f(t))$ for all $t \in \overline{J}$ and defining $A_h : V_h \to V_h$ s.t. $(A_h(v_h), w_h)_L := a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$, (77.7) can be rewritten as follows:

$$\begin{cases} \text{Find } u_h \in C^1(\overline{J}; V_h) \text{ s.t. } u_h(0) = \mathcal{P}_{V_h}(u_0) \text{ and} \\ \partial_t u_h(t) + A_h(u_h(t)) = f_h(t), \qquad \forall t \in \overline{J}. \end{cases} \tag{77.8}$$

Let $\{\varphi_i\}_{i\in\{1:I\}}$ be a basis of V_h with $I := \dim(V_h)$ (the functions $\{\varphi_i\}_{i\in\{1:I\}}$ are usually the global shape functions in V_h). Let $\mathsf{U}(t) \in \mathbb{C}^I$ be the coordinate vector of $u_h(t)$ in this basis for all

$t \in \overline{J}$, i.e., $u_h(t,\boldsymbol{x}) := \sum_{i\in\{1:I\}} \mathsf{U}_i(t)\varphi_i(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$. The *stiffness matrix* $\mathcal{A} \in \mathbb{C}^{I\times I}$ is defined s.t. $\mathcal{A}_{ij} := a_h(\varphi_j,\varphi_i)$, and the *mass matrix* $\mathcal{M} \in \mathbb{C}^{I\times I}$ is defined s.t. $\mathcal{M}_{ij} := (\varphi_j,\varphi_i)_L$, for all $i,j \in \{1:I\}$. The mass matrix is Hermitian positive definite, but the stiffness matrix is in general neither Hermitian nor positive definite. Let $\mathsf{F}(t) := ((f(t),\varphi_i)_L)_{i\in\{1:I\}} \in \mathbb{C}^I$ for all $t \in \overline{J}$, and let U^0 be the coordinate vector of $\mathcal{P}_{V_h}(u_0)$. Then (77.7) is recast as follows: Find $\mathsf{U} \in C^1(\overline{J};\mathbb{C}^I)$ s.t. $\mathsf{U}(0) = \mathsf{U}^0$ and

$$\mathcal{M}\partial_t\mathsf{U}(t) + \mathcal{A}\mathsf{U}(t) = \mathsf{F}(t), \qquad \forall t \in \overline{J}. \tag{77.9}$$

77.2 Implicit Euler scheme

In this section, we use the implicit Euler scheme to approximate in time (77.7), and we perform the stability and error analysis for the fully discrete problem.

77.2.1 Time discrete setting and algebraic realization

As in §67.1, given a positive natural number $N > 0$ we set $\tau := \frac{T}{N}$ and $t_n := n\tau$ for all $n \in \mathcal{N}_\tau := \{1:N\}$. This defines a partition of the time interval $J := (0,T)$ into N subintervals $J_n := (t_{n-1},t_n]$ for all $n \in \mathcal{N}_\tau$. Although this is not a theoretical requirement, we make all these intervals of equal length to simplify the notation. We also assume that the meshes used for the space discretization are time-independent.

The time discretization of (77.7) by the implicit Euler scheme is as follows: First we set $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then letting $f^n := f(t_n) \in L$ for all $n \in \mathcal{N}_\tau$, we obtain $u_h^n \in V_h$ by solving

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a_h(u_h^n, w_h) = \tau(f^n, w_h)_L, \qquad \forall w_h \in V_h. \tag{77.10}$$

Proposition 77.1 (Well-posedness). *Assume that $\tau\Lambda_\flat^- < 1$ (there is no condition on the time step if $\Lambda_\flat \ge 0$). Then* (77.10) *is well-posed.*

Proof. Letting $b(v_h,w_h) := (v_h,w_h)_L + \tau a_h(v_h,w_h)$, the lower bound (77.5) yields

$$\Re\big(b(v_h,v_h)\big) \ge \big(1+\tau\Lambda_\flat\big)\|v_h\|_L^2, \qquad \forall v_h \in V_h, \tag{77.11}$$

so that b_h is L-coercive on V_h if $\tau\Lambda_\flat^- < 1$ (since $\Lambda_\flat \ge -\Lambda_\flat^-$, we have $1+\tau\Lambda_\flat \ge 1-\tau\Lambda_\flat^- > 0$). Therefore, there is a unique $u_h^n \in V_h$ such that $b_h(u_h^n,w_h) = (u_h^{n-1}+\tau f^n, w_h)_L$ for all $w_h \in V_h$. □

Let $\mathsf{U}^n \in \mathbb{C}^I$ be the coordinate vector of u_h^n in the basis $\{\varphi_i\}_{i\in\{1:I\}}$ for all $n \in \overline{\mathcal{N}}_\tau := \{0:N\}$. The algebraic realization of the scheme (77.10) is as follows: For all $n \in \mathcal{N}_\tau$, find $\mathsf{U}^n \in \mathbb{C}^I$ s.t.

$$(\mathcal{M}+\tau\mathcal{A})\mathsf{U}^n = \mathcal{M}\mathsf{U}^{n-1} + \tau\mathsf{F}^n, \tag{77.12}$$

with $\mathsf{F}^n := ((f^n,\varphi_i)_L)_{i\in\{1:I\}} \in \mathbb{C}^I$. The proof of Proposition 77.1 shows that the matrix $\mathcal{M}+\tau\mathcal{A}$ is positive definite (hence invertible) if $\tau\Lambda_\flat^- < 1$.

77.2.2 Stability

We establish in this section a stability estimate for the implicit Euler time-stepping scheme (77.10). To prepare for the error analysis, we consider a variant of the time-stepping scheme (77.10), where

we allow for a slightly more general right-hand side. Specifically, given $u_h^0 \in L$, we obtain $u_h^n \in V_h$ for all $n \in \mathcal{N}_\tau$ by solving

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a_h(u_h^n, w_h) = \tau(\alpha^n, w_h)_L, \tag{77.13}$$

for all $w_h \in V_h$, with $\alpha^n \in L$. The scheme (77.10) is recovered by setting $\alpha^n := f^n$ in (77.13). For any time sequence $\alpha_\tau := (\alpha^n)_{n\in\mathcal{N}_\tau} \in (L)^N$, we set $\|\alpha_\tau\|^2_{\ell^2((0,t_n);L)} := \sum_{m\in\{1:n\}} \tau\|\alpha^m\|_L^2$ for all $n \in \mathcal{N}_\tau$.

Lemma 77.2 (Stability). *Assume that $\tau \le \frac{1}{4}\rho$. Let $u_{h\tau} \in (V_h)^N$ solve (77.13) and assume that α_τ is bounded in the $\|\cdot\|_{\ell^2(J;L)}$-norm. For all $n \in \mathcal{N}_\tau$, we have*

$$\|u_h^n\|_L \le e^{\frac{2t_n}{\rho}}\big(\|u_h^0\|_L + \rho^{\frac12}\|\alpha_\tau\|_{\ell^2((0,t_n);L)}\big). \tag{77.14}$$

Proof. Notice that $\tau \le \frac{1}{4}\rho$ implies that $\tau\Lambda_\flat^- \le \frac{1}{8} < 1$ so that the discrete problem (77.13) is well-posed (see Proposition 77.1). Just like in the space semi-discrete case, we test (77.13) with u_h^n and take the real part of the equation. Using the lower bound (77.5) and the identity $(u_h^n - u_h^{n-1}, u_h^n)_L = \frac12\|u_h^n\|_L^2 - \frac12\|u_h^{n-1}\|_L^2 + \frac12\|u_h^n - u_h^{n-1}\|_L^2$, we obtain

$$\frac12\|u_h^n\|_L^2 + \tau\Lambda_\flat\|u_h^n\|_L^2 + \tau|u_h^n|^2_{\mathcal{MS}} \le \frac12\|u_h^{n-1}\|_L^2 + \tau|(\alpha^n, u_h^n)_L|.$$

Since $|(\alpha^n, u_h^n)_L| \le \|\alpha^n\|_L\|u_h^n\|_L \le \frac{\rho}{2}\|\alpha^n\|_L^2 + \frac{1}{2\rho}\|u_h^n\|_L^2$ and since $\frac{1}{2\rho} - \Lambda_\flat \le \frac{1}{\rho}$, this gives

$$\frac12\|u_h^n\|_L^2 + \tau|u_h^n|^2_{\mathcal{MS}} \le \frac12\|u_h^{n-1}\|_L^2 + \frac{\tau}{\rho}\|u_h^n\|_L^2 + \frac12\rho\tau\|\phi^n\|_L^2.$$

Dropping the nonnegative term $\tau|u_h^n|^2_{\mathcal{MS}}$ from the left-hand side and summing the inequalities for all $l \in \{1:n\}$ gives

$$\|u_h^n\|_L^2 \le \|u_h^0\|_L^2 + \sum_{m\in\{1:n\}} \frac{2\tau}{\rho}\|u_h^m\|_L^2 + \sum_{m\in\{1:n\}} \rho\tau\|\phi^m\|_L^2.$$

We apply the discrete Gronwall lemma from Exercise 68.3 with $\gamma := \frac{2\tau}{\rho} \in (0,1)$ by assumption, $a_m := \|u_h^m\|_L^2$, $b_m := 0$, $c_m := \rho\tau\|\phi^m\|_L^2$, and $B := \|u_h^0\|_L^2$. Since $\gamma \le \frac12$ by assumption, we have $\frac{1}{1-\gamma} \le e^{2\gamma}$, and this yields (77.14) once we take the square root of both sides of the inequality. □

Remark 77.3 (Growth in time). The stability estimate (77.14) allows for an exponential growth on the right-hand side if $\rho \ll T$. This growth can be avoided if $\Lambda_\flat \ge 0$ by replacing ρ by $\rho' := \max(\rho, T)$ in (77.14). □

77.3 Error analysis

In this section, we establish the convergence of the implicit Euler scheme.

77.3.1 Approximation in space

As we did in §76.4.3 for the space semi-discrete problem, there are essentially two ways to perform the analysis. We focus here on the approach that is the most natural in the context of the Hille–Yosida theorem and which relies on the use of the approximation operator $\Pi_h^{\mathrm{A}} : V \to V_h$ s.t.

$$\rho^{-1}(\Pi_h^{\mathrm{A}}(v), w_h)_L + a_h(\Pi_h^{\mathrm{A}}(v), w_h) = \rho^{-1}(v, w_h)_L + (A(v), w_h)_L, \tag{77.15}$$

for all $v \in V$ and all $w_h \in V_h$. This problem is well-posed (recall that the sesquilinear form $\tilde{a}_h(v_h, w_h) := \rho^{-1}(v_h, w_h)_L + a_h(v_h, w_h)$ is L-coercive on V_h with the constant $\mu_0 := \rho^{-1} + \Lambda_\flat \geq \frac{1}{2}\rho^{-1}$). Let us equip the space $V_\sharp$ with the norm

$$\|v\|_{V_\flat}^2 := \rho^{-1}\|v\|_L^2 + |v|_{\mathcal{MS}}^2 + \frac{h}{\beta}\|A_1(v)\|_L^2. \tag{77.16}$$

The approximation properties of Π_h^{A} using this norm are stated in Theorem 58.11 for CIP and in Theorem 59.11 for LPS and SGV. For simplicity, we assume in the rest of this chapter that $h \leq \rho\beta$. The convergence results from Chapters 58-59 with $\mu_0 := \frac{1}{2\rho}$ then imply that for all $v \in H^{k+1}(D;\mathbb{C}^m)$,

$$\|v - \Pi_h^{\mathrm{A}}(v)\|_L \leq \rho^{\frac{1}{2}}\|v - \Pi_h^{\mathrm{A}}(v)\|_{V_\flat} \leq c\,(\rho\beta)^{\frac{1}{2}}h^{k+\frac{1}{2}}|v|_{H^{k+1}(D;\mathbb{C}^m)}. \tag{77.17}$$

77.3.2 Error estimate in the L-norm

Theorem 77.4 ($\ell^\infty(\overline{J};L)$-error estimate). *Let u solve (77.1) and let $u_{h\tau}$ solve (77.10). Assume that $u \in C^2(\overline{J};L) \cap C^1(\overline{J};H^{k+1}(D;\mathbb{C}^m))$. There is c s.t. for all $h \in \mathcal{H} \cap (0,\rho\beta]$, all $\tau \in (0,\frac{1}{4}\rho]$, and all $n \in \mathcal{N}_\tau$,*

$$\|u(t_n) - u_h^n\|_L \leq c\,e^{\frac{2t_n}{\rho}}\Big((t_n\rho)^{\frac{1}{2}}\tau c_1^n(u) + (t_n\beta)^{\frac{1}{2}}h^{k+\frac{1}{2}}c_2^n(u) + (\rho\beta)^{\frac{1}{2}}h^{k+\frac{1}{2}}c_3^n(u)\Big), \tag{77.18}$$

with $c_1^n(u) := |u|_{C^2([0,t_n];L)}$, $c_2^n(u) := \rho|\partial_t u|_{C^0([0,t_n];H^{k+1})} + |u|_{C^0([0,t_n];H^{k+1})}$, and $c_3^n(u) := |u_0|_{H^{k+1}} + |u(t_n)|_{H^{k+1}}$.

Proof. Let us set $e_h^n := u_h^n - \Pi_h^{\mathrm{A}}(u(t_n))$ and $\eta^n := \eta(t_n)$ for all $n \in \overline{\mathcal{N}}_\tau$ with $\eta(t) := \Pi_h^{\mathrm{A}}(u(t)) - u(t)$. Subtracting (77.1) from (77.10) gives for all $w_h \in V_h$,

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau\big(a_h(u_h^n, w_h) - (A(u(t_n)), w_h)_L\big) = -(\eta^n - \eta^{n-1}, w_h)_L - \tau(\psi^n, w_h)_L,$$

with $\psi^n := \frac{1}{\tau}\int_{J_n}(\partial_t u(t) - \partial_t u(t_n))\mathrm{d}t \in L$. Since we have $(A(u(t_n)), w_h)_L = \rho^{-1}(\eta^n, w_h)_L + a_h(\Pi_h^{\mathrm{A}}(u(t_n)), w_h)$ owing to (77.15), rearranging the terms leads to

$$(e_h^n - e_h^{n-1}, w_h)_L + \tau a_h(e_h^n, w_h) = \tau(\alpha^n, w_h)_L, \tag{77.19}$$

with $\alpha^n := -\frac{1}{\tau}(\eta^n - \eta^{n-1}) + \frac{1}{\rho}\eta^n - \psi^n$. Invoking Lemma 77.2 (stability), we infer that

$$\|e_h^n\|_L \leq e^{\frac{2t_n}{\rho}}\big(\|e_h^0\|_L + \rho^{\frac{1}{2}}\|\alpha_\tau\|_{\ell^2((0,t_n);L)}\big),$$

where $\alpha_\tau := (\alpha^n)_{n\in\mathcal{N}_\tau} \in (L)^N$. Invoking the triangle inequality and since $\|e_h^0\|_L \leq \|(I_L - \mathcal{P}_{V_h})(u(0))\|_L + \|\eta^0\|_L \leq 2\|\eta^0\|_L$, we infer that

$$\|u(t_n) - u_h^n\|_L \leq \|\eta^n\|_L + e^{\frac{2t_n}{\rho}}\big(2\|\eta^0\|_L + \rho^{\frac{1}{2}}\|\alpha_\tau\|_{\ell^2((0,t_n);L)}\big).$$

Owing to the approximation property (77.17), we infer that $\|\eta^n\|_L + 2\|\eta^0\|_L \le c(\rho\beta)^{\frac12} h^{k+\frac12} c_3^n(u)$. It remains to bound $\|\alpha_\tau\|_{\ell^2((0,t_n);L)}$. The triangle inequality yields $\|\alpha^n\|_L \le \frac{1}{\tau}\|\eta^n - \eta^{n-1}\|_L + \frac{1}{\rho}\|\eta^n\|_L + \|\psi^n\|_L$. Since A is time-independent, we have $\partial_t \Pi_h^{\mathrm{A}}(u) = \Pi_h^{\mathrm{A}}(\partial_t u)$. This implies that $\eta^n - \eta^{n-1} = \int_{J_n} \partial_t \eta(u(t))\mathrm{d}t = \int_{J_n} \eta(\partial_t u(t))\mathrm{d}t$, so that $\|\eta^n - \eta^{n-1}\|_L \le \tau \|\eta(\partial_t u)\|_{C^0(\overline{J}_n;L)}$. Moreover, we have $\|\eta^n\|_L \le \|\eta(u)\|_{C^0(\overline{J}_n;L)}$. Invoking again the approximation property (77.17), we obtain

$$\tfrac{1}{\tau}\|\eta^n - \eta^{n-1}\|_L + \tfrac{1}{\rho}\|\eta^n\|_L \le c\,\big(\tfrac{\beta}{\rho}\big)^{\frac12} h^{k+\frac12}\big(\rho|\partial_t u|_{C^0(\overline{J}_n;H^{k+1})} + |u|_{C^0(\overline{J}_n;H^{k+1})}\big).$$

Moreover, since $u \in C^2(\overline{J};L)$, we have $\|\psi^n\|_L \le \tau \|\partial_{tt} u\|_{C^2(\overline{J}_n;L)}$. Using that $\|\alpha_\tau\|_{\ell^2((0,t_n);L)} \le t_n^{\frac12} \max_{m\in\{1:n\}} \|\alpha^m\|_L$, the above two bounds give

$$\rho^{\frac12}\|\alpha_\tau\|_{\ell^2((0,t_n);L)} \le c\,(t_n\rho)^{\frac12}\big(\tau c_1^n(u) + \big(\tfrac{\beta}{\rho}\big)^{\frac12} h^{k+\frac12} c_2^n(u)\big). \tag{77.20}$$

The assertion is obtained by putting everything together and using $e^{\frac{2t_n}{\rho}} \ge 1$. This completes the proof. □

77.3.3 Error estimate in the graph norm

Let us set $\|v_\tau\|^2_{\ell^2((0,t_n);V_\flat)} := \sum_{m\in\{1:n\}} \tau\|v^m\|^2_{V_\flat}$ for all $v_\tau := (v^n)_{n\in\mathcal{N}_\tau} \in (V_\sharp)^N$ and all $n \in \mathcal{N}_\tau$. We now derive an error estimate in the $\|\cdot\|_{\ell^2(J;V_\flat)}$-norm. This allows us to gain some control on the error on the spatial derivatives.

Theorem 77.5 ($\ell^2(J;V_\flat)$-error estimate). *Let u solve (77.1) and let $u_{h\tau}$ solve (77.10). Assume that $u \in C^3(\overline{J};L) \cap C^1(\overline{J};H^{k+1}(D;\mathbb{C}^m))$ and $f \in C^2(\overline{J};L)$. Let $e_\tau := (u(t_n) - u_h^n)_{n\in\mathcal{N}_\tau} \in (V_\sharp)^N$. There is c s.t. for all $h \in \mathcal{H} \cap (0,\rho\beta]$, all $\tau \in (0, \frac14\rho]$, and all $n \in \mathcal{N}_\tau$,*

$$\|e_\tau\|_{\ell^2((0,t_n);V_\flat)} \le c\,\big(\tfrac{t_n}{\rho}\big)^{\frac12} e^{\frac{2t_n}{\rho}}\big(\rho\tau\tilde{c}_1^n(u) + (\rho\beta)^{\frac12} h^{k+\frac12}\tilde{c}_2^n(u)\big), \tag{77.21}$$

with $\tilde{c}_1^n(u) := \xi^n c_1^n(u) + (t_n\rho)^{\frac12} c_1^n(A(u))$, $\tilde{c}_2^n(u) := \xi^n c_2^n(u) + (t_n\rho)^{\frac12} c_2^n(\partial_t u)$, $c_1^n(\cdot)$, $c_2^n(\cdot)$ defined in Theorem 77.4, and $\xi^n := \max(\frac{t_n}{\rho}, 1)^{\frac12}$.

Proof. Proceeding as in the proof of Lemma 58.2, one can show that there is $\theta > 0$ such that for all $h \in \mathcal{H}$,

$$\inf_{v_h\in V_h}\sup_{w_h\in V_h} \frac{|a_h(v_h,w_h) + \rho^{-1}(v_h,w_h)_L|}{\|v_h\|_{V_\flat}\|w_h\|_{V_\flat}} \ge \theta > 0. \tag{77.22}$$

The main idea to establish an $\ell^2(J;V_\flat)$-error estimate is to invoke the inf-sup condition (77.22), but to do so we first have to derive an estimate on the time derivative of the error.
(1) Recall that $e_h^n := u_h^n - \Pi_h^{\mathrm{A}}(u(t_n))$ and $\eta^n := \eta(u(t_n))$ for all $n \in \overline{\mathcal{N}}_\tau$ with $\eta(t) := \eta(u(t))$. For any time sequence $(v_h^n)_{n\in\overline{\mathcal{N}}_\tau} \in (V_h)^{N+1}$, we set $D_\tau v_h^n := \frac{1}{\tau}(v_h^n - v_h^{n-1})$ for all $n \in \mathcal{N}_\tau$. We infer from (77.13) that for all $n \in \{2{:}N\}$,

$$(D_\tau u_h^n - D_\tau u_h^{n-1}, w_h)_L + \tau a_h(D_\tau u_h^n, w_h) = \tau(D_\tau f^n, w_h)_L,$$

for all $w_h \in V_h$. Moreover, taking the time derivative of the time-dependent Friedrichs' system leads to

$$(\partial_t u(t_n) - \partial_t u(t_{n-1}), w_h)_L + \tau(A(\partial_t u(t_n)), w_h)_L = \tau(D_\tau f^n + \gamma^n, w_h)_L,$$

with $\gamma^n := \frac{1}{\tau}\int_{J_n}(g(t)-g(t_n))\mathrm{d}t$ and $g := \partial_{tt}u - \partial_t f = A(\partial_t u)$. Let us set $\dot e_h^n := D_\tau u_h^n - \Pi_h^{\mathrm{A}}(\partial_t u(t_n))$ and $\dot\eta^n := \eta(\partial_t u(t_n))$ for all $n \in \mathcal{N}_\tau$. Proceeding as in the proof of Theorem 77.4, we infer that

$$\|\dot e_h^n\|_L \le c\, e^{\frac{2t_n}{\rho}}\Big(\|\dot e_h^1\|_L + (t_n\rho)^{\frac12}\tau c_1^n(A(u)) + (t_n\rho)^{\frac12}\big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}c_2^n(\partial_t u)\Big).$$

Moreover, we have (see Exercise 77.2 for the proof)

$$\|\dot e_h^1\|_L \le c\,\big(\tau c_1^1(u) + \big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}c_2^1(u)\big). \tag{77.23}$$

Letting $\dot e_{h\tau} := (\dot e_h^n)_{n\in\mathcal{N}_\tau}$ and since $\|\dot e_{h\tau}\|_{\ell^2((0,t_n);L)} \le t_n^{\frac12}\max_{m\in\{1:n\}}\|\dot e_h^m\|_L$, the above two bounds imply that

$$\begin{aligned}\rho^{\frac12}\|\dot e_{h\tau}\|_{\ell^2((0,t_n);L)} \le c\,(t_n\rho)^{\frac12}e^{\frac{2t_n}{\rho}}\Big(&\tau\big(c_1^1(u) + (t_n\rho)^{\frac12}c_1^n(A(u))\big)\\ &+ \big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}\big(c_2^1(u) + (t_n\rho)^{\frac12}c_2^n(\partial_t u)\big)\Big).\end{aligned}$$

(2) Combining the inf-sup condition (77.22) with (77.19) yields for all $n \in \mathcal{N}_\tau$,

$$\theta\,\|e_h^n\|_{V_\flat} \le \rho^{\frac12}\big(\|\alpha^n\|_L + \|\dot e_h^n\|_L\big) + \rho^{-\frac12}\|e_h^n\|_L,$$

which implies that $\|e_{h\tau}\|_{\ell^2((0,t_n);V_\flat)} \le c\rho^{\frac12}(\|\alpha_\tau\|_{\ell^2((0,t_n);L)} + \|\dot e_{h\tau}\|_{\ell^2((0,t_n);L)}) + \rho^{-\frac12}\|e_{h\tau}\|_{\ell^2((0,t_n);L)}$. We use the bound (77.20) on $\rho^{\frac12}\|\alpha_\tau\|_{\ell^2((0,t_n);L)}$ already established in the proof of Theorem 77.4 and the bound on $\rho^{\frac12}\|\dot e_{h\tau}\|_{\ell^2((0,t_n);L)}$ established in Step (1). Since $c_1^1(u) \le c_1^n(u)$ and $c_2^1(u) \le c_2^n(u)$, this gives

$$\begin{aligned}\|e_{h\tau}\|_{\ell^2((0,t_n);V_\flat)} \le c\,(t_n\rho)^{\frac12}e^{\frac{2t_n}{\rho}}\Big(&\tau\big(c_1^n(u) + (t_n\rho)^{\frac12}c_1^n(A(u))\big)\\ &+ \big(\tfrac{\beta}{\rho}\big)^{\frac12}h^{k+\frac12}\big(c_2^n(u) + (t_n\rho)^{\frac12}c_2^n(\partial_t u)\big)\Big) + \big(\tfrac{t_n}{\rho}\big)^{\frac12}\max_{m\in\{1:n\}}\|e_h^m\|_L.\end{aligned}$$

Finally, $\|e_h^m\|_L$ is bounded using Theorem 77.4. The assertion follows from the triangle inequality and the approximation property (77.17). □

Remark 77.6 (Literature). The material is adapted from Johnson et al. [201], Guermond [149], Burman and Fernández [66]. The higher-order discretization by means of the discontinuous Galerkin method in time is analyzed in Ern and Schieweck [122]. Space-time discontinuous Galerkin methods have been studied by Monk and Richter [228]. □

Remark 77.7 (Singular perturbation). The above theory can be adapted to solve the time-dependent version of problems like the advection-diffusion problem investigated in Chapter 61, that is, $\partial_t u + B(u) = f$ with $B := A + D : X \subsetneq V_0 \to X'$, where $A : V_0 \to L$ is a maximal monotone (first-order) differential operator and $D : X \to X'$ is a coercive second-order differential operator with $X \subset V_0 \subset L \equiv L' \subset V_0' \subset X'$ (think of $X := H_0^1(D)$ and $X' = H^{-1}(D)$). The coercivity of D (with $\xi := 1$) implies that the evolution equation is parabolic in the sense of Definition 65.3. Let $\lambda_\flat := \inf_{v\in X}\langle D(v), w\rangle_{X',X}/\|v\|_X$. Let $\{X_h\}_{h\in\mathcal{H}}$ be some X-conforming approximation sequence for X. The approximation theory from Chapter 67 to Chapter 70 fully applies if the Peclet number $\mathrm{Pe} := \frac{\beta h}{\lambda_\flat}$ is $\mathcal{O}(1)$ (see Chapter 61 and (61.8)). But the coercivity is not strong enough to guarantee that the Galerkin approximation is satisfactory if $\mathrm{Pe} \gg 1$, since one essentially has $B \approx A$ in this case. The stabilization theory presented above can be applied to this problem. The time-stepping can be done, e.g., using the implicit Runge–Kutta (IRK) schemes from §69.2.4 and §70.1.3, see Exercise 77.3. We refer the reader to Guermond [149], Burman and Fernández [66], Burman and Ern [64] for further developments; see also Exercise 77.1. □

Exercises

Exercise 77.1 (Implicit advection-diffusion). Consider the 1D equation $\mu\partial_t u+\beta\partial_x u-\nu\partial_{xx}u = f$ in $D := (0,1)$, $t>0$, where $\mu\in\mathbb{R}_+$, $\beta\in\mathbb{R}$, $\nu\in\mathbb{R}_+$, $f\in L^2(D)$, boundary conditions $u(0)=0$, $u(1)=0$, and initial data $u_0=0$. Let $\mathcal{T}_h$ be the mesh composed of the cells $[ih,(i+1)h]$, $i\in\{0{:}I\}$, with uniform meshsize $h := \frac{1}{I+1}$. Let $V_h := P^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions that are zero at 0 and at 1 (see (19.37)). Let $(\varphi_i)_{i\in\{1:I\}}$ be the global Lagrange shape functions associated with the nodes $x_i := ih$ for all $i\in\{1{:}I\}$. (i) Write the fully discrete version of the problem in V_h using the implicit Euler time-stepping scheme. Denote the time step by τ and the discrete time nodes by $t_n := n\tau$ for all $n\in\mathcal{N}_\tau$. (ii) Prove a stability estimate. (*Hint*: consider the test function $2\tau u_h^n$ and introduce the Poincaré–Steklov constant C_{PS} s.t. $C_{\mathrm{PS}}\|v\|_{L^2(D)}\le \ell_D\|\partial_x v\|_{L^2(D)}$ for all $v\in H^1_0(D)$.) (iii) Letting $u_h^n := \sum_{i\in\{1:I\}}\mathsf{U}_i^n\varphi_i$ and $\mathsf{F}_i := \frac{1}{h}\int_D f\varphi_i\,\mathrm{d}x$ for all $i\in\{1{:}I\}$, write the linear system solved by the vector $\mathsf{U}^n := (\mathsf{U}_i^n)_{i\in\{1:I\}}$. (iv) Prove that $\max_{i\in\{1:I\}}\mathsf{U}_i^n \le \frac{\tau}{\mu}\max_{i\in\{1:I\}}\mathsf{F}_i + \max_{i\in\{1:I\}}\mathsf{U}_i^{n-1}$ if $\nu>|\beta|h$ and $\tau\ge\frac{\mu h^2}{3(2\nu-|\beta|h)}$. (*Hint*: consider the index $j\in\{1{:}I\}$ s.t. $\mathsf{U}_j^n=\max_{i\in\{1:I\}}\mathsf{U}_i^n$.)

Exercise 77.2 (Bound on $\|\dot e_h^1\|_L$). Prove (77.23). (*Hint*: use that $e_h^0=0$ and test (77.19) with $n:=1$ against $w_h := e_h^1$.)

Exercise 77.3 (IRK for advection-diffusion). Consider the advection-diffusion problem from Remark 77.7. Write the time-stepping process in functional and algebraic form using the IRK formalism from §69.2.4 and §70.1.3.

Exercise 77.4 (Implicit Euler, analysis using $\mathcal{P}_{V_h}$). The objective of this exercise is to derive an $\ell^\infty(\overline{J};L)$-error estimate for the implicit Euler scheme by using the operator $\mathcal{P}_{V_h}$ instead of the operator Π_h^{A} as was done in §77.3. We assume that $\tau\le\frac14\rho$. (i) Consider the following scheme: Given $u_h^0\in L$, one obtains $u_h^n\in V_h$ for all $n\in\mathcal{N}_\tau$ by solving

$$(u_h^n-u_h^{n-1},w_h)_L+\tau a_h(u_h^n,w_h)=\tau\phi^n(w_h),\qquad \forall w_h\in V_h,$$

with $\phi^n\in V_h'$. Set $\phi_\tau := (\phi^n)_{n\in\mathcal{N}_\tau}\in(V_h')^N$ and $\|\phi_\tau\|^2_{\ell^2((0,t_n);V_{h\flat}')} := \sum_{m\in\{1:n\}}\tau\|\phi^m\|^2_{V_{h\flat}'}$ with $\|\phi^m\|_{V_{h\flat}'} := \sup_{w_h\in V_h}\frac{|\phi^m(w_h)|}{\|w_h\|_{V_\flat}}$ and the norm $\|\cdot\|_{V_\flat}$ is defined as $\|v\|_{V_\flat} := \rho^{-1}\|v\|_L^2+\|v\|_{\mathcal{MS}}$ (this is the definition used in the proof of Theorem 76.19; it differs from (77.16)). Show that for all $n\in\mathcal{N}_\tau$,

$$\|u_h^n\|_L\le e^{\frac{2t_n}{\rho}}\big(\|u_h^0\|_L+\|\phi_\tau\|_{\ell^2((0,t_n);V_{h\flat}')}\big).$$

(*Hint*: adapt the proof of Lemma 77.2.) (ii) Let $e_h^n := u_h^n-\mathcal{P}_{V_h}(u(t_n))$ and $\eta^n := \mathcal{P}_{V_h}(u(t_n))-u(t_n)$ for all $n\in\overline{\mathcal{N}}_\tau$. Prove that $(e_h^n-e_h^{n-1},w_h)_L+\tau a_h(e_h^n,w_h) = -\tau\phi^n(w_h)$ for all $w_h\in V_h$, with $\phi^n\in V_h'$ s.t.

$$\begin{aligned}\phi^n(w_h) &= (\psi^n+K(\eta^n)-\mathcal{X}\eta^n,w_h)_L+s_h(\mathcal{P}_{V_h}(u(t_n)),w_h)\\ &\quad+\frac12((\mathcal{M}^{\mathrm{BP}}+\mathcal{N})\eta^n,w_h)_{L(\partial D)}-(\eta^n,A_1(w_h))_L,\end{aligned}$$

and $\psi^n := \frac{1}{\tau}\int_{J_n}(\partial_t u(t)-\partial_t u(t_n))\mathrm{d}t\in L$. (*Hint*: see (76.27).) (iii) Let u solve (77.1) and let $u_{h\tau}$ solve (77.10). Assume that $u\in C^2(\overline{J};L)\cap C^0(\overline{J};H^{k+1}(D;\mathbb{C}^m))$. Prove that there is c s.t. for all $h\in\mathcal{H}$, all $\tau>0$, and all $n\in\mathcal{N}_\tau$,

$$\|u(t_n)-u_h^n\|_L\le c\,e^{\frac{2t_n}{\rho}}\Big(\tau(\rho t_n)^{\frac12}c_1(t_n;u)+\big(h^{\frac12}+\big(\tfrac{t_n}{\rho}\big)^{\frac12}\max(\rho\beta,h)^{\frac12}h^{k+\frac12}\big)c_2(t_n;u)\Big),$$

with $c_1(t_n;u) := \|\partial_{tt}u\|_{C^0([0,t_n];L)}$ and $c_2(t_n;u) := |u|_{C^0([0,t_n];H^{k+1}(D;\mathbb{C}^m))}$. (*Hint*: see the proof of Theorem 76.19 and use Step (i).)

Chapter 78

Explicit time discretization

In this chapter, we continue our investigation on the time approximation of the semi-discrete problem (77.8), i.e.,

$$\begin{cases} \text{Find } u_h \in C^1(\overline{J}; V_h) \text{ s.t. } u_h(0) = \mathcal{P}_{V_h}(u_0) \text{ and} \\ \partial_t u_h(t) + A_h(u_h(t)) = f_h(t), \qquad \forall t \in \overline{J}, \end{cases} \tag{78.1}$$

where the setting for the space discretization is described in §77.1. We first discuss generic properties of *explicit Runge–Kutta* schemes (ERK). Then we analyze the explicit Euler scheme, second-order two-stage ERK schemes, and third-order three-stage ERK schemes. The key advantage of explicit schemes over implicit schemes is that the linear algebra at each time step is greatly simplified since one has to invert only the mass matrix. However, the stability of ERK schemes requires that the time step τ be limited by a *CFL condition* of the form $\tau \le \lambda_0 \tau_\gamma(h)$ with the time scale $\tau_\gamma(h) := \left(\frac{h}{\beta}\right)^\gamma \rho^{1-\gamma}$ for some real numbers $\lambda_0 > 0$ and $\gamma \ge 1$ (the time scale ρ is defined in (77.3)). For $\gamma := 1$, this condition takes the usual form $\tau \le \lambda_0 \frac{h}{\beta}$. The acronym CFL stands for Courant–Friedrichs–Levy. The nondimensional number $\frac{\tau\beta}{h}$ has been introduced in the context of the approximation of the wave equation in [91, §II.2, p. 61] (see also [92, §II.2, p. 228] for the English translation).

78.1 Explicit Runge–Kutta (ERK) schemes

We start by reviewing generic properties of explicit Runge–Kutta schemes.

78.1.1 Butcher tableau

Just like IRK methods introduced in §69.2.4 and §70.1.3, s-stage ERK methods, $s \ge 1$, are characterized by their Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:s\}}$, $\{b_i\}_{i\in\{1:s\}}$, $\{c_i\}_{i\in\{1:s\}}$. Recall that the time discretization of (78.1) by any RK scheme, whether implicit or explicit, is as follows. One first sets $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then for all $n \in \mathcal{N}_\tau$ one sets $t_{n,j} := t_{n-1} + c_j\tau$ for all $j \in \{1:s\}$ and seeks $\{u_h^{n,i}\}_{i\in\{1:s\}} \subset V_h$ by solving the following system of equations:

$$u_h^{n,i} - u_h^{n-1} = \tau \sum_{j\in\{1:s\}} a_{ij}\big(f_h(t_{n,j}) - A_h(u_h^{n,j})\big), \qquad \forall i \in \{1:s\}. \tag{78.2}$$

Finally, the update at the time t_n is obtained by setting

$$u_h^n := u_h^{n-1} + \tau \sum_{j\in\{1:s\}} b_j\big(f_h(t_{n,j}) - A_h(u_h^{n,j})\big). \tag{78.3}$$

The key difference between implicit and explicit RK schemes is that the Butcher matrix a of an explicit scheme is strictly lower triangular, so that the *Butcher tableau* becomes (compare with (69.21))

$$\begin{array}{c|cccc} c_1 & 0 & & & \\ c_2 & a_{21} & 0 & & \\ \vdots & \vdots & & \ddots & \\ c_s & a_{s1} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & \cdots & b_{s-1} & b_s \end{array} \tag{78.4}$$

As a result, we have $u_h^{n,1} - u_h^{n-1} = 0$, and for all $i \ge 2$ the summation in (78.2) can be restricted to $j \in \{1{:}i{-}1\}$. Hence, $u_h^{n,i}$ can be explicitly evaluated in terms of the previously computed values $\{u_h^{n,j}\}_{j\in\{1:i-1\}}$ for all $i \in \{2{:}s\}$.

Recalling the mass matrix $\mathcal{M} \in \mathbb{C}^{I\times I}$ and the stiffness matrix $\mathcal{A} \in \mathbb{C}^{I\times I}$ introduced in §77.1.2, the algebraic realization of (78.1) is $\mathcal{M}\partial_t \mathsf{U}(t) + \mathcal{A}\mathsf{U}(t) = \mathsf{F}(t)$ (see (77.9)), which we rewrite as

$$\partial_t \mathsf{U}(t) = \tilde{\mathcal{A}}\mathsf{U}(t) + \tilde{\mathsf{F}}(t), \quad \tilde{\mathcal{A}} := -\mathcal{M}^{-1}\mathcal{A}, \quad \tilde{\mathsf{F}}(t) := \mathcal{M}^{-1}\mathsf{F}(t). \tag{78.5}$$

The algebraic realization of (78.2)-(78.3) is then

$$\mathsf{U}^{n,i} - \mathsf{U}^{n-1} = \tau \sum_{j\in\{1:s\}} a_{ij}\big(\tilde{\mathcal{A}}\mathsf{U}^{n,j} + \tilde{\mathsf{F}}^{n,j}\big), \quad \forall i \in \{1{:}s\}, \tag{78.6a}$$

$$\mathsf{U}^{n} := \mathsf{U}^{n-1} + \tau \sum_{j\in\{1:s\}} b_j\big(\tilde{\mathcal{A}}\mathsf{U}^{n,j} + \tilde{\mathsf{F}}^{n,j}\big). \tag{78.6b}$$

where $\mathsf{F}^{n,j} \in \mathbb{C}^I$ is the coordinate vector of $f_h(t_{n,j})$ in the basis $\{\varphi_i\}_{i\in\{1:I\}}$ of V_h, $\mathsf{U}^{n,j} \in \mathbb{C}^I$ that of $u_h^{n,j}$, and $\mathsf{U}^n \in \mathbb{C}^I$ that of u_h^n.

An equivalent way to proceed, which is often used in the literature, consists of introducing the discrete time derivative $\mathsf{K}^{n,i} := \tilde{\mathcal{A}}(\mathsf{U}^{n,i}) + \tilde{\mathsf{F}}(t_{n,i})$ for all $i \in \{1{:}s\}$. Then one proceeds as follows at each time step for ERK schemes. One first sets $\mathsf{U}^{n,1} := \mathsf{U}^{n-1}$ and $\mathsf{K}^{n,1} := \tilde{\mathcal{A}}\mathsf{U}^{n-1} + \tilde{\mathsf{F}}(t_{n,1})$, then for all $i \ge 2$ one computes

$$\mathsf{K}^{n,i} := \tilde{\mathcal{A}}\Big(\mathsf{U}^{n-1} + \tau \sum_{j\in\{1:i-1\}} a_{ij}\mathsf{K}^{n,j}\Big) + \tilde{\mathsf{F}}^{n,i}, \tag{78.7}$$

and the update at time t_n is $\mathsf{U}^n := \mathsf{U}^{n-1} + \tau\sum_{j\in\{1:s\}} b_j\mathsf{K}^{n,j}$.

78.1.2 Examples

Examples of ERK schemes are the first-order one-stage Euler method, the second-order two-stage Heun scheme, the *midpoint rule*, and the third-order three-stage Heun scheme. The Butcher tableaux for these four methods are shown here from left to right, respectively:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{3} & \frac{1}{3} & 0 & \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array} \tag{78.8}$$

Example 78.1 (Explicit Euler). One computes the discrete time derivative $\mathsf{K}^{n,1} := \tilde{\mathcal{A}}\mathsf{U}^{n-1} + \mathsf{F}^{n-1}$ with $t_{n,1} := t_{n-1}$ and sets $\mathsf{U}^n := \mathsf{U}^{n-1} + \tau\mathsf{K}^{n,1}$. □

Example 78.2 (Second-order Heun). One computes the discrete time derivatives $\mathsf{K}_1 := \tilde{\mathcal{A}}\mathsf{U}^{n-1} + \tilde{\mathsf{F}}^{n,1}$ and $\mathsf{K}_2 := \tilde{\mathcal{A}}(\mathsf{U}^{n-1} + \tau\mathsf{K}_1) + \tilde{\mathsf{F}}^{n,2}$ with $t_{n,1} := t_{n-1}$ and $t_{n,2} := t_n$, and one sets $\mathsf{U}^n := \mathsf{U}^{n-1} + \frac{1}{2}\tau(\mathsf{K}_1 + \mathsf{K}_2)$. □

Example 78.3 (Midpoint rule). One computes the discrete time derivatives $\mathsf{K}_1 := \tilde{\mathcal{A}}\mathsf{U}^{n-1} + \tilde{\mathsf{F}}^{n,1}$ and $\mathsf{K}_2 := \tilde{\mathcal{A}}(\mathsf{U}^{n-1} + \frac{1}{2}\tau\mathsf{K}_1) + \tilde{\mathsf{F}}^{n,2}$ with $t_{n,1} := t_{n-1}$ and $t_{n,2} := t_{n-1} + \frac{1}{2}\tau$, and one sets $\mathsf{U}^n := \mathsf{U}^{n-1} + \tau\mathsf{K}_2$. □

Example 78.4 (Third-order Heun). One computes the discrete velocities $\mathsf{K}_1 := \tilde{\mathcal{A}}\mathsf{U}^{n-1} + \tilde{\mathsf{F}}^{n,1}$, $\mathsf{K}_2 := \tilde{\mathcal{A}}(\mathsf{U}^{n-1} + \frac{1}{3}\tau\mathsf{K}_1) + \tilde{\mathsf{F}}^{n,2}$, $\mathsf{K}_3 := \tilde{\mathcal{A}}(\mathsf{U}^{n-1} + \frac{2}{3}\tau\mathsf{K}_2) + \tilde{\mathsf{F}}^{n,3}$ with $t_{n,1} := t_{n-1}$, $t_{n,2} := t_{n-1} + \frac{1}{3}\tau$, and $t_{n,3} := t_{n-1} + \frac{2}{3}\tau$, and one sets $\mathsf{U}^n := \mathsf{U}^{n-1} + \frac{1}{4}\tau(\mathsf{K}_1 + 3\mathsf{K}_3)$. □

78.1.3 Order conditions

Let $\partial_t\mathsf{U}(t) = \mathsf{L}(t, \mathsf{U}(t))$ be some nonlinear ODE system in $\mathbb{C}^I$. For instance, we have $\mathsf{L}(t, \mathsf{V}) := \tilde{\mathcal{A}}\mathsf{V} + \tilde{\mathsf{F}}(t)$ for the semi-discrete time-dependent Friedrichs' system. Let $n \in \mathcal{N}_\tau$, set $\mathsf{U}^{n-1} := \mathsf{U}(t_{n-1})$, and consider any RK scheme (whether implicit or explicit) to step from U^{n-1} to U^n. Assuming that $\mathsf{U} \in C^\infty(\overline{J}; \mathbb{C}^I)$, we call *truncation error* at t_n of the RK scheme the quantity $\mathsf{U}(t_n) - \mathsf{U}^n$. If the local truncation error is $\mathcal{O}(\tau^{p+1})$ in some norm, then the method is said to be of order p (here the precise definition of the norm does not matter since $\mathbb{C}^I$ is finite-dimensional and, for the time being, we are not concerned about the size of the constant multiplying τ^{p+1}). The Butcher coefficients must satisfy algebraic relations to guarantee the order of the method. The following order conditions were established in [77, Thm. 7] and are often called *Butcher's simplifying assumptions* in the ODE literature; see also Hairer et al. [176, §II.7, Thm. 7.4], [175, §IV.5,Thm. 5.1], and Exercises 70.3 and 78.2.

Theorem 78.5 (Butcher). *Consider an s-stage RK method with Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:s\}}$, $\{b_j\}_{j\in\{1:s\}}$, $\{c_j\}_{j\in\{1:s\}}$. A sufficient condition for the method to be of order $p \geq 1$ is that there exist $\eta, \zeta \in \mathbb{N}$ with $p \leq \min(2(1+\eta), 1+\eta+\zeta)$ such that the following is satisfied:*

$$\sum_{j\in\{1:s\}} b_j c_j^{q-1} = \frac{1}{q}, \qquad \forall q \in \{1{:}p\}, \tag{78.9a}$$

$$\sum_{j\in\{1:s\}} a_{ij} c_j^{q-1} = \frac{c_i^q}{q}, \qquad \forall i \in \{1{:}s\},\ \forall q \in \{1{:}\eta\}, \tag{78.9b}$$

$$\sum_{i\in\{1:s\}} b_i c_i^{q-1} a_{ij} = \frac{b_j}{q}(1 - c_j^q), \quad \forall j \in \{1{:}s\},\ \forall q \in \{1{:}\zeta\}. \tag{78.9c}$$

By convention, (78.9b) *is empty if $\eta = 0$ and* (78.9c) *is empty if $\zeta = 0$.*

The reader is invited to verify that $p = 1$, $\eta = 1$, $\zeta = 0$ for both the explicit and the implicit Euler schemes, $p = 2$, $\eta = 1$, $\zeta = 1$ for the two-stage Heun scheme, $p = 2$, $\eta = 1$, $\zeta = 0$ for the midpoint rule, and $p = 3$, $\eta = 1$, $\zeta = 0$ for three-stage Heun scheme (notice that $p > 1 + \eta + \zeta$ in this case).

Remark 78.6 (Coefficients $\{c_i\}_{i\in\{1:s\}}$). Whenever $\eta \geq 1$ (78.9b) with $q := 1$ gives $c_i = \sum_{j\in\{1:s\}} a_{ij}$ for all $i \in \{1{:}s\}$. This implies that $c_1 = 0$ for ERK schemes s.t. $\eta \geq 1$ in (78.9b). Note that $a_{11} := 0$, $b_1 := 1$ with any $c_1 \in (0, 1]$ is a legitimate 1-stage ERK scheme for which

$\eta = 0$. For simplicity, and as usually done in the literature, we henceforth assume that $c_i \in [0,1]$ for all $i \in \{1\!:\!s\}$, and we take $c_1 := 0$ unless stated otherwise. □

Lemma 78.7 (Necessary order conditions). *Consider an s-stage RK method with Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:s\}}$, $\{b_j\}_{j\in\{1:s\}}$, $\{c_j\}_{j\in\{1:s\}}$. A necessary condition for the scheme to be of order $p \geq 1$ is*

$$\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \dots a_{j_{r-1} j_r} c_{j_r}^{q-1} = \frac{(q-1)!}{(q-1+r)!}, \tag{78.10}$$

for all $r \in \{1\!:\!p\}$ and all $q \in \{1\!:\!p{-}r{+}1\}$, where we use the convention $a_{j_1 j_2} \dots a_{j_{r-1} j_r} := 1$ if $r = 1$. Moreover, when applied to any linear system of the form $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$, the condition (78.10) *is also sufficient to guarantee that the scheme is of order p.*

Proof. (1) By linearity, U solves $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$ with $\mathsf{U}(t_n) = \mathsf{U}^{n-1}$ iff $\mathsf{U} = \mathsf{U}_1 + \mathsf{U}_2$ where U_1 solves $\partial_t \mathsf{U}_1 = \tilde{\mathcal{A}}\mathsf{U}_1$ with $\mathsf{U}_1(t_{n-1}) = \mathsf{U}^{n-1}$ and $\partial_t \mathsf{U}_2 = \tilde{\mathcal{A}}\mathsf{U}_2 + \tilde{\mathsf{F}}$ with $\mathsf{U}_2(t_{n-1}) = 0$. Hence, an RK scheme is of order p for the solution of $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$ with $\mathsf{U}(t_{n-1}) = \mathsf{U}^{n-1}$ iff it is of order p for the solution of $\partial_t \mathsf{U}_1 = \tilde{\mathcal{A}}\mathsf{U}_1$ with $\mathsf{U}_1(t_{n-1}) = \mathsf{U}^{n-1}$ and it is of order p for the solution of $\partial_t \mathsf{U}_2 = \tilde{\mathcal{A}}\mathsf{U}_2 + \tilde{\mathsf{F}}$ with $\mathsf{U}_2(t_{n-1}) = 0$.
(2) Let us prove that the condition (78.10) with $q := 1$ is necessary and sufficient to have a scheme of order p for the linear system $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U}$, i.e., letting U^n be the update produced by the RK scheme, we need to show that $\mathsf{U}^n = \sum_{r\in\{0:p\}} \frac{\tau^r}{r!} \tilde{\mathcal{A}}^r \mathsf{U}^{n-1} + \mathcal{O}(\tau^{p+1})$ iff

$$\sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \dots a_{j_{r-1} j_r} = \frac{1}{r!}, \qquad \forall r \in \{1\!:\!p\}. \tag{78.11}$$

To prove this claim, we replace the values of $\mathsf{U}^{n,j}$ repeatedly p times. Since $\tilde{\mathcal{A}}$ does not depend on time and the matrix-vector multiplication is linear, this process gives

$$\begin{aligned}
\mathsf{U}^n &= \mathsf{U}^{n-1} + \tau \sum_{j_1\in\{1:s\}} b_{j_1} \tilde{\mathcal{A}} \mathsf{U}^{n,j_1} \\
&= \mathsf{U}^{n-1} + \tau \sum_{j_1\in\{1:s\}} b_{j_1} \tilde{\mathcal{A}} \mathsf{U}^{n-1} + \tau^2 \sum_{j_1,j_2\in\{1:s\}} b_{j_1} a_{j_1,j_2} \tilde{\mathcal{A}}^2 \mathsf{U}^{n,j_2} \\
&= \mathsf{U}^{n-1} + \sum_{r\in\{1:p\}} \tau^r \sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1} a_{j_1 j_2} \dots a_{j_{r-1} j_r} \tilde{\mathcal{A}}^r \mathsf{U}^{n-1} \\
&\quad + \tau^{p+1} \sum_{j_1,\dots,j_{p+1}\in\{1:s\}} b_{j_1} a_{j_1 j_2} \dots a_{j_p j_{p+1}} \tilde{\mathcal{A}}^{p+1} \mathsf{U}^{n,j_{p+1}}.
\end{aligned} \tag{78.12}$$

The last term in the above identity is $\mathcal{O}(\tau^{p+1})$. Hence, (78.11) are necessary conditions to get the order p (they are also sufficient for the linear ODE under consideration). Notice in passing that the term $\mathcal{O}(\tau^{p+1})$ is zero for explicit methods with $s = p$ stages (see Remark 78.10(i)).
(3) It remains to consider the system $\partial_t \mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$ with $\mathsf{U}(t_{n-1}) = 0$. The Taylor expansion of $\mathsf{U}(t_n)$ up to the order p gives $\mathsf{U}(t_n) = \tau \mathsf{G}_p(t_{n-1}) + \mathcal{O}(\tau^{p+1})$ with

$$\mathsf{G}_p(t) := \sum_{r\in\{1:p\}} \frac{\tau^{r-1}}{r!} \sum_{q\in\{1:r\}} \tilde{\mathcal{A}}^{r-q} \partial_t^{q-1} \tilde{\mathsf{F}}(t). \tag{78.13}$$

The above sum can be reorganized as follows:

$$\mathsf{G}_p(t) = \sum_{r\in\{1:p\}} (\tau\tilde{\mathcal{A}})^{r-1} \sum_{q\in\{1:p-r+1\}} \frac{\tau^{q-1}}{(q-1+r)!} \partial_t^{q-1} \tilde{\mathsf{F}}(t).$$

Moreover, proceeding as in Exercise 78.1(ii) by successive elimination of the intermediate stages, we infer that the RK scheme leads to $\mathsf{U}^n = \tau\tilde{\mathsf{G}}_p^n + \mathcal{O}(\tau^{p+1})$ with

$$\tilde{\mathsf{G}}_p^n := \sum_{r\in\{1:p\}} (\tau\tilde{\mathcal{A}})^{r-1} \sum_{j_1,\dots,j_r\in\{1:s\}} b_{j_1}a_{j_1j_2}\dots a_{j_{r-1}j_r}\tilde{\mathsf{F}}^{n,j_r}. \tag{78.14}$$

Writing the Taylor expansion of $\tilde{\mathsf{F}}^{n,j_r}$ up to the order $(p-r)$ and equating the coefficients with $\mathsf{G}_p(t_{n-1})$ leads to the assertion. □

Remark 78.8 (Theorem 78.5 vs. Lemma 78.7). (78.9a) is equivalent to (78.10) with $r := 1$. Moreover, it is shown in Exercise 78.2 that if (78.9a) is met and (78.9b)–(78.9c) hold with $p \le \eta + \zeta + 1$, then (78.10) holds for all $r \in \{1{:}p\}$ and all $q \in \{1{:}p{-}r{+}1\}$. □

Remark 78.9 (Minimal consistency requirement). The minimal consistency requirement to have a first-order method is obtained by taking $p := 1$ in (78.10). Then $r = 1$, $q = 1$, and this gives $\sum_{j\in\{1:s\}} b_j = 1$. □

Remark 78.10 (Consequences). (i) A consequence of (78.12) is that any ERK scheme of order $p \ge 1$ with $s = p$ stages is such that $\mathsf{U}^n = \sum_{r\in\{0:p\}} \frac{\tau^r}{r!}\tilde{\mathcal{A}}^r\mathsf{U}^{n-1}$. Indeed, the $\mathcal{O}(\tau^{p+1})$ remainder in (78.12) is necessarily zero (the only way a term in the sum is nonzero is if $j_1 > j_2 \dots > j_{p+1}$, but these inequalities cannot be satisfied since we have $p+1$ indices in $\{1{:}s\}$ and $s = p$). (ii) (78.10) with $r := 1$ (or equivalently (78.9a)) is a necessary and sufficient condition to guarantee that an RK scheme is of order p for the uncoupled ODE system $\partial_t\mathsf{U} = \tilde{\mathsf{F}}$. Indeed, in this case, the only nonzero contribution to $\tilde{\mathsf{G}}_p^n$ in (78.14) is obtained for $r := 1$; see also Exercise 78.1(ii). □

Example 78.11 (ERK schemes, $p \in \{2,3\}$). Consider a p-th-order ERK scheme with $s = p$ stages. For $p := 2$, the relations are $b_1 + b_2 = 1$, $b_1c_1 + b_2c_2 = \frac{1}{2}$, and $b_2a_{21} = \frac{1}{2}$. For $p := 3$, the relations are $b_1+b_2+b_3 = 1$, $b_1c_1+b_2c_2+b_3c_3 = \frac{1}{2}$, $b_1c_1^2+b_2c_2^2+b_3c_3^2 = \frac{1}{3}$, $b_2a_{21}+b_3a_{31}+b_3a_{32} = \frac{1}{2}$, $b_2a_{21}c_1 + b_3a_{31}c_1 + b_3a_{32}c_2 = \frac{1}{6}$, and $b_3a_{32}a_{21} = \frac{1}{6}$. The reader is invited to verify that these identities hold true for the second-order Heun scheme, the midpoint rule, and the third-order Heun scheme. □

78.2 Explicit Euler scheme

The explicit Euler scheme is defined by the leftmost Butcher tableau in (78.8). It is the simplest explicit Runge–Kutta method. First we set $u_h^0 := \mathcal{P}_{V_h}(u_0)$, then we obtain $u_h^n \in V_h$ for all $n \in \mathcal{N}_\tau$ by solving

$$(u_h^n - u_h^{n-1}, w_h)_L + \tau a_h(u_h^{n-1}, w_h) = \tau(\alpha^{n,1}, w_h)_L, \tag{78.15}$$

for all $w_h \in V_h$, with $\alpha^{n,1} := f(t_{n,1}) := f^{n-1} \in L$ (since $t_{n,1} := t_{n-1}$ for the explicit Euler scheme). The algebraic realization of (78.15) is

$$\mathcal{M}\mathsf{U}^n = (\mathcal{M} - \tau\mathcal{A})\mathsf{U}^{n-1} + \tau\mathsf{F}^{n-1}, \tag{78.16}$$

which only requires to invert the mass matrix at each time step (compare with (77.12)). The difficulty with the explicit Euler scheme is that its stability requires a rather stringent condition on the time step. More precisely, setting $\tau_2(h) := \left(\frac{h}{\beta}\right)^2\rho^{-1}$, we introduce a positive (nondimensional) number λ_0 and say that the *2-CFL condition* is satisfied whenever $\tau \le \lambda_0\tau_2(h)$. Since all the stability constants are going to be increasing functions of λ_0, one should in practice pick $\lambda_0 = \mathcal{O}(1)$. We perform the stability analysis using generic functions $\alpha^{n,1} \in L$. We set $\alpha_\tau := (\alpha^{n,1})_{n\in\mathcal{N}_\tau} \in (L)^N$ and consider the norm $\|\alpha_\tau\|^2_{\ell^2((0,t_n);L)} := \sum_{m\in\{1:n\}} \tau\|\alpha^{m,1}\|_L^2$.

Lemma 78.12 (Stability). *Let $\alpha_\tau \in (L)^N$ and let $u_{h\tau} \in (V_h)^N$ solve (78.15). For every $\lambda_0 > 0$, there are c_1, c_2 (depending on λ_0) s.t. the following holds for all $h \in \mathcal{H} \cap (0, \rho\beta]$, all $\tau \in (0, \lambda_0 \tau_2(h)]$, and all $n \in \mathcal{N}_\tau$,*

$$\|u_h^n\|_L^2 \le e^{c_1 \frac{t_n}{\rho}} \Big(\|u_h^0\|_L^2 + c_2 \rho \|\alpha_\tau\|^2_{\ell^2((0,t_n);L)} \Big). \tag{78.17}$$

Proof. We use the symbols c, c_1, c_2 to denote generic constants that may depend on λ_0 but are uniform w.r.t. τ and h and whose value can change at each occurrence. Consider the test function $w_h := u_h^{n-1}$ in (78.15), take the real part, use the identity $(u_h^n - u_h^{n-1}, u_h^{n-1})_L = \frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 - \frac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2$ and the lower bound (77.5) on the sesquilinear form a_h. Dropping the nonnegative term $\tau |u_h^{n-1}|^2_{\mathcal{MS}}$ from the left-hand side and rearranging the terms, this gives

$$\frac{1}{2}\|u_h^n\|_L^2 - \frac{1}{2}\|u_h^{n-1}\|_L^2 \le -\tau \Lambda_\flat \|u_h^{n-1}\|_L^2 + \tau |(\alpha^{n,1}, u_h^{n-1})_L| + \frac{1}{2}\|u_h^n - u_h^{n-1}\|_L^2.$$

Since $|(\alpha^{n,1}, u_h^{n-1})_L| \le \|\alpha^{n,1}\|_L \|u_h^{n-1}\|_L \le \frac{\rho}{2}\|\alpha^{n,1}\|_L^2 + \frac{1}{2\rho}\|u_h^{n-1}\|_L^2$ and since $\frac{1}{2\rho} - \Lambda_\flat \le \frac{1}{\rho}$, we infer that

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le \frac{2\tau}{\rho}\|u_h^{n-1}\|_L^2 + \rho\tau\|\alpha^{n,1}\|_L^2 + \|u_h^n - u_h^{n-1}\|_L^2.$$

The novelty with respect to the implicit Euler scheme is the positive term $\|u_h^n - u_h^{n-1}\|_L^2$ which we need to bound. This is done by using the CFL restriction on the time step. Invoking an inverse inequality and the bound $|w_h|_{\mathcal{MS}} \le c\beta^{\frac{1}{2}} h^{-\frac{1}{2}} \|w_h\|_L$, using that $h \le \rho\beta$ by assumption, we infer that $|a_h(v_h, w_h)| \le c\beta h^{-1} \|v_h\|_L \|w_h\|_L$ for all $v_h, w_h \in V_h$, where c depends on the constant $\check{c}$ defined in (77.6). Using (78.15), we obtain

$$|(u_h^n - u_h^{n-1}, w_h)_L| \le c\,\tau \tfrac{\beta}{h} \|u_h^{n-1}\|_L \|w_h\|_L + \tau \|\alpha^{n,1}\|_L \|w_h\|_L.$$

Hence, we have

$$\|u_h^n - u_h^{n-1}\|_L^2 \le c_1 \tau^2 \tfrac{\beta^2}{h^2} \|u_h^{n-1}\|_L^2 + c_2 \tau^2 \|\alpha^{n,1}\|_L^2.$$

For the first term on the right-hand side, we invoke the 2-CFL condition which implies that $\tau^2 \frac{\beta^2}{h^2} \le \lambda_0 \frac{\tau}{\rho}$. For the second term, we use the bound $\tau \le \lambda_0 \rho$, which results from the 2-CFL condition and the assumption $h \le \beta\rho$. This gives $\|u_h^n - u_h^{n-1}\|_L^2 \le c_1 \frac{\tau}{\rho} \|u_h^{n-1}\|_L^2 + c_2 \rho\tau \|\alpha^{n,1}\|_L^2$. Putting everything together, we obtain

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le c_1 \frac{\tau}{\rho} \|u_h^{n-1}\|_L^2 + c_2 \rho\tau \|\alpha^{n,1}\|_L^2.$$

We conclude by induction using that $1 + \gamma \le e^\gamma$ with $\gamma := c_1 \frac{\tau}{\rho}$. □

Remark 78.13 (CFL condition, error estimate). In general, the CFL condition $\tau \le \lambda_0 \tau_2(h)$ is too stringent to be useful in practice since the upper bound is quadratic with respect to the meshsize. Nevertheless, assuming that the solution to (76.6) is smooth enough, it is possible to proceed as in §78.3 and show that there are c, c' s.t. for all $h \in \mathcal{H} \cap (0, \rho\beta]$, all $\tau \in (0, \lambda_0 \tau_2(h)]$, and all $n \in \mathcal{N}_\tau$,

$$\|u(t_n) - u_h^n\|_L \le c' e^{c \frac{t_n}{\rho}} \Big(\tau (t_n \rho)^{\frac{1}{2}} c_1^n(u) + \big(\max(t_n, \rho)\beta\big)^{\frac{1}{2}} h^{k+\frac{1}{2}} c_2^n(u) \Big),$$

with $c_1^n(u) := \|\partial_t^2 u\|_{C^0([0,t_n];L)}$ and $c_2^n(u) := |u|_{C^0([0,t_n];H^{k+1})}$. The same estimate can be obtained for any Butcher coefficient $c_1 \in [0, 1]$. □

Remark 78.14 (Variants). The stability of the explicit Euler scheme can be obtained under the usual CFL condition $\tau \le \lambda_0 \frac{h}{\beta}$ if some first-order linear stabilization is introduced. In this case, the accuracy in space reduces to $\mathcal{O}(h)$ at best (see Exercise 78.4). It is shown in Bonito et al. [38] that high-order accuracy in space can be preserved under the usual CFL condition if the stabilization is nonlinear, i.e., the dependence on u_h^{n-1} is nonlinear. Moreover, it is shown in Exercise 78.5 that the explicit Euler scheme with mass lumping and without linear stabilization is unconditionally unstable, i.e., no time step restriction can make the method stable in any reasonable sense. □

78.3 Second-order two-stage ERK schemes

The goal of this section is to establish an $\ell^\infty(\overline{J};L)$-error estimate for second-order *two-stage ERK* schemes. Since there are many such schemes, we are going to study the stability of one representative scheme and then show that the error analysis is valid for all second-order two-stage ERK schemes. The representative scheme we have in mind is as follows: Setting as usual $u_h^0 := \mathcal{P}_{V_h}(u_0)$, one builds two sequences $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ and $y_{h\tau} := (y_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$ so that for all $n \in \mathcal{N}_\tau$ and all $w_h \in V_h$,

$$(y_h^n - u_h^{n-1}, w_h)_L + \tau a_h(u_h^{n-1}, w_h) = \tau(\alpha^{n,1}, w_h)_L, \tag{78.18a}$$

$$(u_h^n - \tfrac12(y_h^n + u_h^{n-1}), w_h)_L + \tfrac12\tau a_h(y_h^n, w_h) = \tfrac12\tau(\alpha^{n,2}, w_h)_L, \tag{78.18b}$$

with $\alpha^{n,1} := f^{n-1}$ and $\alpha^{n,2} := f^{n-1} + \tau\partial_t f^{n-1}$. Notice that (78.18) is not an ERK scheme since the right-hand side of (78.18b) requires the evaluation of $\partial_t f$, whereas ERK schemes only sample values of f in $\overline{J}_n$ (see Remark 78.6). The reason we specifically consider the scheme (78.18) is that eliminating y_h^n gives

$$u_h^n = (I_{V_h} - \tau A_h + \tfrac12\tau^2 A_h^2)(u_h^{n-1}) + \tau \mathsf{G}_2(t_{n-1}), \tag{78.19}$$

with $\mathsf{G}_2(t_{n-1}) := f_h^{n-1} + \frac12\tau\partial_t f_h^{n-1} - \frac12\tau A_h(f_h^{n-1})$. In other words, u_h^n exactly reproduces the second-order Taylor expansion of the semi-discrete solution $u_h(t)$ at t_n (see (78.30) and (78.13) with $p := 2$).

Lemma 78.15 (ERK schemes, $p = 2$). *Consider a second-order two-stage ERK scheme defined by its Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:2\}}$, $\{b_i\}_{i\in\{1:2\}}$, $\{c_i\}_{i\in\{1:2\}}$. Let $u_{h\tau}$ be the sequence approximating* (78.1) *that is produced by this second-order two-stage ERK scheme. For all $n \in \mathcal{N}_\tau$, set*

$$r_h^{n,2} := \mathcal{P}_{V_h}(2b_1 f(t_{n,1}) + 2b_2 f(t_{n,2}) - 2f^{n-1} - \tau\partial_t f^{n-1}). \tag{78.20}$$

The following holds true: (i) *$u_{h\tau}$ is also the sequence produced by* (78.18) *with the data $\alpha^{n,2}$ replaced by $\widetilde{\alpha}^{n,2} := \alpha^{n,2} + r_h^{n,2}$ in* (78.18b). (ii) *There is c that only depends on $\{b_i\}_{i\in\{1:2\}}$ s.t.*

$$\|r_h^{n,2}\|_L \le c\,\tau^2 \|\partial_t^2 f\|_{C^0(\overline{J}_n;L)}. \tag{78.21}$$

Proof. (i) Let $\widetilde{u}_{h\tau}$ be the sequence produced by (78.18) with $\alpha^{n,2}$ replaced by $\widetilde{\alpha}^{n,2} := \alpha^{n,2} + r_h^{n,2}$. Owing to (78.19), we have

$$\widetilde{u}_h^n = (I_{V_h} - \tau A_h + \tfrac12\tau^2 A_h^2)(\widetilde{u}_h^{n-1}) + \tau\big(\mathsf{G}_2(t_{n-1}) + \tfrac12 r_h^{n,2}\big).$$

Moreover, eliminating the intermediate stage in the ERK scheme leads to

$$u_h^n = (I_{V_h} - \tau A_h + \tfrac12\tau^2 A_h^2)(u_h^{n-1}) + \tau\big(b_1 f_h(t_{n,1}) + b_2 f_h(t_{n,2}) - \tfrac12\tau A_h(f_h^{n-1})\big),$$

where we used $b_1+b_2=1$, $b_2a_{21}=\frac12$ (see Lemma 78.7). Recalling the definition of $\mathsf{G}_2(t_{n-1})$, an induction argument shows that the sequences $u_{h\tau}$ and $\widetilde{u}_{h\tau}$ coincide if $r_h^{n,2}=\mathcal{P}_{V_h}(r^{n,2})$ with $r^{n,2}\in L$ such that

$$\tfrac12 r^{n,2}:=b_1f(t_{n,1})+b_2f(t_{n,2})-f^{n-1}-\tfrac12\tau\partial_tf^{n-1}.$$

(ii) Since $\|r_h^{n,2}\|_L\le\|r^{n,2}\|_L$, it suffices to bound $r^{n,2}$. Using that $b_1+b_2=1$, $b_1c_1+b_2c_2=\frac12$ (notice that altogether we used the three necessary order conditions from Lemma 78.7 for $p:=2$), and $f(t_{n,j})-f^{n-1}-c_j\tau\partial_tf^{n-1}=\int_{t_{n-1}}^{t_{n,j}}(t_{n,j}-t)\partial_{tt}f(t)\,\mathrm{d}t$ for all $j\in\{1\!:\!2\}$ gives

$$\tfrac12 r^{n,2}=\sum_{j\in\{1:2\}}b_j\int_{t_{n-1}}^{t_{n,j}}(t_{n,j}-t)\partial_{tt}f(t)\,\mathrm{d}t.$$

Hence, (78.21) is satisfied with c only depending on $\{b_i\}_{i\in\{1:2\}}$. □

Example 78.16 (Heun scheme and midpoint rule). For the second-order Heun scheme, we have $r^{n,2}:=f^n-f^{n-1}-\tau\partial_tf^{n-1}$, i.e., $\mathcal{P}_{V_h}(\widetilde{\alpha}^{n,2}):=f_h^n$, whereas for the midpoint rule we have $r^{n,2}:=2(f(t_{n-1}+\frac12\tau)-f^{n-1})-\tau\partial_tf^{n-1}$, i.e., $\mathcal{P}_{V_h}(\widetilde{\alpha}^{n,2}):=2f_h(t_{n-1}+\frac12\tau)-f_h^{n-1}$. □

The scheme (78.18) turns out to be conditionally stable. More precisely, setting $\tau_{4/3}(h):=\big(\frac{h}{\beta}\big)^{\frac43}\rho^{-\frac13}$, the stability analysis will reveal that the stability constants depend on the ratio $\tau/\tau_{4/3}(h)$. To account for this phenomenon, we introduce a positive (nondimensional) number λ_0 and say that the pair (τ,h) satisfy the *4/3-CFL condition* if

$$\tau\le\lambda_0\tau_{4/3}(h). \tag{78.22}$$

Since all the stability constants are going to be increasing functions on λ_0, one should in practice pick $\lambda_0=\mathcal{O}(1)$. Notice that $\tau\frac{\beta}{h}\le\big(\frac{\tau}{\rho}\big)^{\frac14}\lambda_0^{\frac34}$ when $\tau\le\lambda_0\tau_{4/3}(h)$, and that $\tau_{4/3}(h)\le\rho$ when $h\le\rho\beta$. Hence, $\tau\le\lambda_0\rho$ when $h\le\rho\beta$ and $\tau\le\lambda_0\tau_{4/3}(h)$.

Lemma 78.17 (Stability, 4/3-CFL condition). *Let $\alpha_\tau^1:=(\alpha^{1,n})_{n\in\mathcal{N}_\tau}$, $\alpha_\tau^2:=(\alpha^{2,n})_{n\in\mathcal{N}_\tau}$, both in $(L)^N$, and let $u_{h\tau}\in(V_h)^N$ solve (78.18). For every $\lambda_0>0$, there are c_1,c_2 (depending on λ_0) s.t. the following holds true for all $h\in\mathcal{H}\cap(0,\rho\beta]$, all $\tau\in(0,\lambda_0\tau_{4/3}(h)]$, and all $n\in\mathcal{N}_\tau$,*

$$\|u_h^n\|_L^2\le e^{c_1\frac{t_n}{\rho}}\Big(\|u_h^0\|_L^2+c_2\rho\big(\|\alpha_\tau^1\|_{\ell^2((0,t_n);L)}^2+\|\alpha_\tau^2\|_{\ell^2((0,t_n);L)}^2\big)\Big). \tag{78.23}$$

Proof. We use the symbols c, c_1, c_2 to denote generic constants that may depend on λ_0 but are uniform w.r.t. τ and h and whose value can change at each occurrence. Taking $w_h:=u_h^{n-1}$ in (78.18a) and $v_h:=2y_h^n$ in (78.18b), adding the two equations, taking the real part, using the lower bound (77.5) on the sesquilinear form a_h, and rearranging the terms, we infer that

$$\begin{aligned}\|u_h^n\|_L^2-\|u_h^{n-1}\|_L^2+\tau|u_h^{n-1}|_{\mathcal{MS}}^2+\tau|y_h^n|_{\mathcal{MS}}^2&\le\|u_h^n-y_h^n\|_L^2\\&\quad+\tau\Big(-\Lambda_\flat\|u_h^{n-1}\|_L^2+|(\alpha^{n,1},u_h^{n-1})_L|-\Lambda_\flat\|y_h^n\|_L^2+|(\alpha^{n,2},y_h^n)_L|\Big).\end{aligned}$$

Using the Cauchy–Schwarz inequality and since $\frac12\rho^{-1}-\Lambda_\flat\le\rho^{-1}$, we obtain

$$\begin{aligned}\|u_h^n\|_L^2-\|u_h^{n-1}\|_L^2+\frac12\tau|u_h^{n-1}|_{\mathcal{MS}}^2+\frac12\tau|y_h^n|_{\mathcal{MS}}^2&\le\|u_h^n-y_h^n\|_L^2\\&\quad+\frac{\tau}{\rho}\big(\|u_h^{n-1}\|_L^2+\|y_h^n\|_L^2\big)+\frac{\rho\tau}{2}\big(\|\alpha^{n,1}\|_L^2+\|\alpha^{n,2}\|_L^2\big).\end{aligned}$$

We still need to bound $\|u_h^n - y_h^n\|_L^2$. Combining the two equations in (78.18) yields

$$(u_h^n - y_h^n, w_h)_L = \tfrac{1}{2}\tau\big(\alpha^{n,2} - \alpha^{n,1}, w_h\big)_L - \tfrac{1}{2}\tau a_h(y_h^n - u_h^{n-1}, w_h),$$

for all $w_h \in V_h$. Let us denote by $\mathfrak{T}_1(w_h), \mathfrak{T}_2(w_h)$ the two terms on the right-hand side. The Cauchy–Schwarz and the triangle inequality yield

$$|\mathfrak{T}_1(w_h)| \le \tfrac{1}{2}\tau(\|\alpha^{n,1}\|_L + \|\alpha^{n,2}\|_L)\|w_h\|_L.$$

Moreover, invoking an inverse inequality and since $|w_h|_{\mathcal{MS}} \le c\,\beta^{\frac12}h^{-\frac12}\|w_h\|_L$, we have $|a_h(v_h, w_h)| \le c\beta h^{-1}\|v_h\|_L\|w_h\|_L$ for all $v_h, w_h \in V_h$. This implies that

$$|\mathfrak{T}_2(w_h)| \le c\,\tau\tfrac{\beta}{h}\|y_h^n - u_h^{n-1}\|_L\|w_h\|_L.$$

Combining the bounds on $\mathfrak{T}_1(w_h)$ and $\mathfrak{T}_2(w_h)$ we infer that

$$\|u_h^n - y_h\|_L^2 \le c_1\tau^2\tfrac{\beta^2}{h^2}\|y_h^n - u_h^{n-1}\|_L^2 + c_2\tau^2(\|\alpha^{n,1}\|_L^2 + \|\alpha^{n,2}\|_L^2).$$

Similar arguments using (78.18a) imply that

$$\|y_h^n - u_h^{n-1}\|_L^2 \le c_1\tau^2\tfrac{\beta^2}{h^2}\|u_h^{n-1}\|_L^2 + c_2\tau^2\|\alpha^{n,1}\|_L^2. \tag{78.24}$$

Using that $\tau\frac{\beta}{h} \le \lambda_0$, the above two bounds give

$$\|u_h^n - y_h^n\|_L^2 \le c_1\tau^4\tfrac{\beta^4}{h^4}\|u_h^{n-1}\|_L^2 + c_2\tau^2(\|\alpha^{n,1}\|_L^2 + \|\alpha^{n,2}\|_L^2).$$

We can now invoke the 4/3-CFL condition (78.22) which yields $\tau^4\frac{\beta^4}{h^4} \le \frac{\tau}{\rho}\lambda_0^3$. Putting everything together, recalling that our assumptions imply $\tau \le \lambda_0\rho$, and dropping the seminorms on the left-hand side we obtain

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le c_1\frac{\tau}{\rho}\big(\|u_h^{n-1}\|_L^2 + \|y_h^n\|_L^2\big) + c_2\tau\rho\big(\|\alpha^{n,1}\|_L^2 + \|\alpha^{n,2}\|_L^2\big).$$

Since $\|y_h^n\|_L \le \|u_h^{n-1}\|_L + \|y_h^n - u_h^{n-1}\|_L \le c_1\|u_h^{n-1}\|_L + c_2\tau^2\|\alpha^{n,1}\|_L$ (owing to (78.24) and $\tau\frac{\beta}{h} \le \lambda_0^{\frac34}$), and using $\tau \le \lambda_0\rho$, we obtain

$$\|u_h^n\|_L^2 - \|u_h^{n-1}\|_L^2 \le c_1\frac{\tau}{\rho}\|u_h^{n-1}\|_L^2 + c_2\tau\rho\big(\|\alpha^{n,1}\|_L^2 + \|\alpha^{n,2}\|_L^2\big).$$

We conclude by induction using that $1 + \gamma \le e^\gamma$ with $\gamma := c_1\frac{\tau}{\rho}$. □

We can now derive an error estimate in the $\ell^\infty(\overline{J}; L)$-norm for any second-order two-stage ERK scheme.

Theorem 78.18 ($\ell^\infty(\overline{J}; L)$-error estimate). *Let u solve (76.6). Assume $u \in C^3(\overline{J}; L) \cap C^1(\overline{J}; H^{k+1}(D; \mathbb{C}^m))$ and $f \in C^2(\overline{J}; L)$. Let $u_{h\tau}$ be given by any second-order two-stage ERK scheme. For every $\lambda_0 > 0$, there are c, c' s.t. for all $h \in \mathcal{H} \cap (0, \rho\beta]$, all $\tau \in (0, \lambda_0\tau_{4/3}(h)]$, and all $n \in \mathcal{N}_\tau$,*

$$\begin{aligned}\|u(t_n) - u_h^n\|_L \le c'\,e^{c\frac{t_n}{\rho}}\Big(&\tau^2(t_n\rho)^{\frac12}\big(c_1^n(u) + d_1^n(f)\big)\\ &+ (t_n\beta)^{\frac12}h^{k+\frac12}c_2^n(u) + (\rho\beta)^{\frac12}h^{k+\frac12}c_3^n(u)\Big),\end{aligned} \tag{78.25}$$

with $c_1^n(u) := \|\partial_t^3 u\|_{C^0([0,t_n];L)}$, $c_2^n(u) := \sum_{q\in\{0:1\}}\rho^q|\partial_t^q u|_{C^0([0,t_n];H^{k+1})}$, $c_3^n(u) = |u(t_n)|_{H^{k+1}}$, and $d_1^n := \|\partial_t^2 f\|_{C^0([0,t_n];L)}$.

Proof. (1) We draw the attention of the reader on two points concerning the way the error equations are constructed. First, since we only want to estimate the error on the end-of-stage update u_h^n and not that on the intermediate stages of the ERK scheme, we invoke Lemma 78.15 and consider that the sequence $u_{h\tau}$ is produced by the scheme (78.18) with $\alpha^{n,1} := f^{n-1}$ and $\alpha^{n,2} := f^{n-1} + \tau\partial_t f^{n-1} + r_h^{n,2}$ where $r_h^{n,2} \in V_h$ is defined in (78.20) and satisfies $\|r_h^{n,2}\|_L \le c\tau^2 \|\partial_{tt} f\|_{C^0(\overline{J}_n;L)}$ (see (78.21)). Second we realize that if we are not careful and write the error equation for the first step (78.18a) like we would for the explicit or implicit forward Euler scheme, the consistency error would scale like $\mathcal{O}(\tau)$, and the global error would then be $\mathcal{O}(\tau)$ owing to (78.23). To avoid this difficulty, we define $y(t) := u(t) + \tau\partial_t u(t)$, and we compare u_h^n with $\Pi_h^{\mathrm{A}}(u(t_n))$ and y_h^n with $\Pi_h^{\mathrm{A}}(y(t_n))$ for all $n \in \mathcal{N}_\tau$, where Π_h^{A} is the approximation operator introduced in (77.15). We set

$$\begin{aligned} e_h^n &:= u_h^n - \Pi_h^{\mathrm{A}}(u(t_n)), \qquad & \eta^n &:= \Pi_h^{\mathrm{A}}(u(t_n)) - u(t_n), \\ z_h^n &:= y_h^n - \Pi_h^{\mathrm{A}}(y(t_n)), \qquad & \zeta^n &:= \Pi_h^{\mathrm{A}}(y(t_n)) - y(t_n). \end{aligned}$$

(2) We are now ready to derive the error equations. We observe that $y(t_{n-1}) - u(t_{n-1}) = \tau\partial_t u(t_{n-1})$ which gives

$$(y(t_{n-1}) - u(t_{n-1}), w_h)_L + \tau(A(u(t_{n-1})), w_h)_L = \tau(f^{n-1}, w_h)_L,$$

for all $w_h \in V_h$. Subtracting this equation from (78.18a) gives

$$(z_h^n - e_h^{n-1}, w_h)_L + \tau\big(a_h(u_h^{n-1}, w_h) - (A(u(t_{n-1})), w_h)_L\big) = (\eta^{n-1} - \eta^n, w_h)_L.$$

Since the definition of Π_h^{A} implies that $(A(u(t_{n-1})), w_h)_L) = \rho^{-1}(\eta^{n-1}, w_h)_L + a_h(\Pi_h^{\mathrm{A}}(u(t_{n-1})), w_h)_L$, rearranging the terms we obtain

$$(z_h^n - e_h^{n-1}, w_h)_L + \tau a_h(e_h^{n-1}, w_h) = \tau(\beta^{n,1}, w_h)_L,$$

with $\beta^{n,1} := -\frac{1}{\tau}(\eta^n - \eta^{n-1}) + \frac{1}{\rho}\eta^{n-1}$. Moreover, since $\partial_t y(t_{n-1}) + A(y(t_{n-1})) = f(t_{n-1}) + \tau f(t_{n-1})$, and $u(t_n) - \frac{1}{2}(y(t_n) + u(t_{n-1})) = \tau\partial_t y(t_{n-1}) + \frac{\tau}{2}\psi^n$, with $\psi^n := \frac{1}{\tau}\int_{J_n}(t_n - t)^2\partial_{ttt}u(t)\mathrm{d}t$, a direct calculation (see Exercise 78.6) shows that for all $w_h \in V_h$,

$$\begin{aligned} (u(t_n) - \tfrac{1}{2}(y(t_{n-1}) + u(t_{n-1})), w_h)_L + \tfrac{1}{2}\tau(A(y(t_{n-1})), w_h)_L \\ = \tfrac{1}{2}\tau\big((\psi^n, w_h)_L + (f^{n-1} + \tau\partial_t f^{n-1}, w_h)_L\big). \end{aligned} \tag{78.26}$$

Subtracting this equation from (78.18b) with $\alpha^{n,2} := f^{n-1} + \tau\partial_t f^{n-1} + r_h^{n,2}$, and reasoning as above gives

$$(e_h^n - \tfrac{1}{2}(z_h^n + e_h^{n-1}), w_h)_L + \tfrac{1}{2}\tau a_h(z_h^n, w_h) = \tfrac{1}{2}\tau(\beta^{n,2}, w_h)_L,$$

with $\beta^{n,2} := \frac{1}{\rho}\zeta^n - \frac{1}{\tau}(\eta^n - \frac{1}{2}(\zeta^n + \eta^{n-1})) + \frac{1}{2}r_h^{n,2} - \frac{1}{2}\psi^n$.

(3) We now invoke the approximation property (77.17) of Π_h^{A}. The inequality $\|\beta^{n,1}\|_L \le \frac{1}{\tau}\|\eta^n - \eta^{n-1}\|_L + \frac{1}{\rho}\|\eta^n\|_L$ implies that

$$\|\beta^{n,1}\|_L \le c\,\big(\tfrac{\beta}{\rho}\big)^{\frac{1}{2}} h^{k+\frac{1}{2}} \chi_2^n(u),$$

with $\chi_2^n(u) := |u|_{C^0(\overline{J}_n;H^{k+1})} + \rho|\partial_t u|_{C^0(\overline{J}_n;H^{k+1})}$. We proceed similarly to bound $\|\beta^{n,2}\|_L$. Using that $\zeta(t) = \eta(t) + \tau\partial_t\eta(t)$ for all $t \in \overline{J}$, we obtain $\frac{1}{\rho}\|\zeta^n\|_L \le c\big(\frac{\beta}{\rho}\big)^{\frac{1}{2}} h^{k+\frac{1}{2}}(|u|_{C^0(\overline{J}_n;H^{k+1})} + \tau|\partial_t u|_{C^0(\overline{J}_n;H^{k+1})}) \le c\big(\frac{\beta}{\rho}\big)^{\frac{1}{2}} h^{k+\frac{1}{2}}\chi_2^n(u)$ since our assumptions imply that $\tau \le \lambda_0\rho$. Moreover, since $\eta^n - \frac{1}{2}(\zeta^n + \eta^{n-1}) = \frac{1}{2}(\eta^n - \eta^{n-1}) - \frac{\tau}{2}\partial_t\eta^n$ and $\tau \le \lambda_0\rho$, we infer that

$$\|\tfrac{1}{\tau}(\eta^n - \tfrac{1}{2}(\zeta^n + \eta^{n-1}))\|_L \le c\,\rho\big(\tfrac{\beta}{\rho}\big)^{\frac{1}{2}} h^{k+\frac{1}{2}} |\partial_t u|_{C^0(\overline{J}_n;H^{k+1})}.$$

Finally, setting $\chi_1^n(u,f) := \|\partial_t^3 u\|_{C^0(\overline{J}_n;L)} + \|\partial_t^2 f\|_{C^0(\overline{J}_n;L)}$, we have $\|r_h^{n,2} + \psi^n\|_L \le c\tau^2\chi_1^n(u,f)$. Altogether this yields

$$\|\beta^{n,2}\|_L \le c\left(\tau^2\chi_1^n(u,f) + \left(\tfrac{\beta}{\rho}\right)^{\frac12} h^{k+\frac12}\chi_2^n(u)\right).$$

The error estimate follows from Lemma 78.17 with $\beta^{n,i}$ in lieu of $\alpha^{n,i}$ for all $i \in \{1{:}2\}$, the triangle inequality, the approximation property (77.17), $\max_{m\in\{1:n\}}\chi_1^m(u,f) \le c_1^n(u) + d_1^n(f)$, and $\max_{m\in\{1:n\}}\chi_2^m(u) = c_2^n(u)$. □

Remark 78.19 (Literature). The convergence for linear H^1-conforming finite elements in the scalar case was analyzed in Ying [288] under the 4/3-CFL condition (78.22). The general case concerning the order of the spatial discretization was treated in Zhang and Shu [292] for scalar equations discretized using discontinuous finite elements and in Burman et al. [74] for Friedrichs' systems discretized using either stabilized H^1-conforming or discontinuous finite elements. The material in this section is based on [74, Lem. 3.2 & Thm. 3.1]. Moreover, it is shown in [292, 74] that in the case of linear elements, i.e., $k = 1$, the same stability and convergence results hold true under the usual CFL condition $\tau\frac{\beta}{h} \le \lambda_0$ with $\lambda_0 > 0$ small enough; see, e.g., [74, Thm. 3.2]. Finally, the $\ell^\infty(\overline{J};L)$-error estimate can also be established by deriving the error equations using the L-orthogonal projection $\mathcal{P}_{V_h}$ instead of Π_h^{A} (recall that $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is quasi-uniform). In this case, the regularity assumption on the solution is $u \in C^3(\overline{J};L)\bigcap_{q=0}^1 C^q(\overline{J};H^{k+1-q}(D;\mathbb{C}^m))$. □

Remark 78.20 (4/3-CFL condition). The 4/3-CFL condition (78.22) is not very restrictive when used with finite elements of degree $k \ge 2$. Indeed, since the RK2 scheme is second order in time, the time discretization error converges essentially as $\mathcal{O}(h^{\frac83})$, whereas the space discretization error converges as $\mathcal{O}(h^{k+\frac12})$. Thus, if $k = 2$, both sources of error are almost equilibrated asymptotically, whereas for $k \ge 3$, a stronger restriction on the time step is needed to equilibrate the time and space errors. □

78.4 Third-order three-stage ERK schemes

The convergence analysis for *third-order three-stage ERK* schemes proceeds as for second-order two-stage ERK schemes. The representative scheme we consider is as follows: Setting as usual $u_h^0 := \mathcal{P}_{V_h}(u_0)$, one builds three sequences $u_{h\tau} := (u_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$, $y_{h\tau} := (y_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$, and $z_{h\tau} := (z_h^n)_{n\in\mathcal{N}_\tau} \in (V_h)^N$, so that the following holds true for all $n \in \mathcal{N}_\tau$ and all $w_h \in V_h$:

$$(y_h^n - u_h^{n-1}, w_h)_L + \tau a_h(u_h^{n-1}, w_h) = \tau(\alpha^{n,1}, w_h)_L, \tag{78.27a}$$

$$(z_h^n - \tfrac12(y_h^n + u_h^{n-1}), w_h)_L + \tfrac12\tau a_h(y_h^n, w_h) = \tfrac12\tau(\alpha^{n,2}, w_h)_L, \tag{78.27b}$$

$$(u_h^n - \tfrac13(z_h^n + y_h^n + u_h^{n-1}), w_h)_L + \tfrac13\tau a_h(z_h^n, w_h) = \tfrac13\tau(\alpha^{n,3}, w_h)_L, \tag{78.27c}$$

with $\alpha^{n,1} := f^{n-1}$, $\alpha^{n,2} := f^{n-1} + \tau\partial_t f^{n-1}$, and $\alpha^{n,3} := f^{n-1} + \tau\partial_t f^{n-1} + \frac12\tau^2\partial_{tt} f^{n-1}$.

Lemma 78.21 (ERK schemes, $p = 3$). *Consider a third-order three-stage ERK scheme defined by its Butcher coefficients $\{a_{ij}\}_{i,j\in\{1:3\}}$, $\{b_i\}_{i\in\{1:3\}}$, $\{c_i\}_{i\in\{1:3\}}$. Let $u_{h\tau}$ be the sequence approximating* (78.1) *that is produced by this third-order three-stage ERK scheme. For all $n \in \mathcal{N}_\tau$, set $r_h^{n,3} := 3\big(\mathcal{P}_{V_h}(r_1^{n,3}) - \tau A_h(\mathcal{P}_{V_h}(r_2^{n,3}))\big)$ with*

$$r_1^{n,3} := b_1 f(t_{n,1}) + b_2 f(t_{n,2}) + b_3 f(t_{n,3}) - f^{n-1} - \tfrac12\tau\partial_t f^{n-1} - \tfrac16\tau^2\partial_{tt} f^{n-1},$$

$$r_2^{n,3} := (b_2 a_{21} + b_3 a_{31}) f(t_{n,1}) + b_3 a_{32} f(t_{n,2}) - \tfrac12 f^{n-1} - \tfrac16\tau\partial_t f^{n-1}.$$

The following holds true: (i) $u_{h\tau}$ *is also the sequence produced by* (78.27) *with the data* $\alpha^{n,3}$ *replaced by* $\widetilde{\alpha}^{n,3} := \alpha^{n,3} + r_h^{n,3}$ *in* (78.27c). (ii) *There is* c *that only depends on* $\{a_{ij}\}_{i,j\in\{1:3\}}$, $\{b_i\}_{i\in\{1:3\}}$ *s.t.*

$$\|r_h^{n,3}\|_L \le c\,\tau^3\big(\|\partial_t^3 f\|_{C^0(\overline{J}_n;L)} + \frac{1}{\rho}\|\partial_t^2 f\|_{C^0(\overline{J}_n;H^1)}\big). \tag{78.28}$$

Proof. See Exercise 78.7. □

The main difference between RK2 and RK3 is that the stability for the scheme (78.27) can be established under the usual *CFL condition* $\tau \le \lambda_0\tau_1(h)$ with $\tau_1(h) = \frac{h}{\beta}$, i.e., $\tau \le \lambda_0\frac{h}{\beta}$, provided λ_0 is chosen small enough (see Burman et al. [74, Eq. (4.18)]). For brevity, we skip the stability result (see [74, Lem. 4.3]) and just state the $\ell^\infty(\overline{J};L)$-error estimate.

Theorem 78.22 ($\ell^\infty(\overline{J};L)$-error estimate). *Let* u *solve* (76.6). *Assume* $u \in C^4(\overline{J};L)\cap C^2(\overline{J};H^{k+1}(D;\mathbb{C}^m))$ *and* $f \in C^3(\overline{J};L)\cap C^2(\overline{J};H^1(D;\mathbb{C}^m))$. *Let* $u_{h\tau}$ *be given by any third-order three-stage ERK scheme. There exists* $\lambda_0 > 0$ *and there are* c, c', *s.t. for all* $h \in \mathcal{H}\cap(0,\rho\beta]$, *all* $\tau \in (0, \lambda_0\tau_1(h))]$, *and all* $n \in \mathcal{N}_\tau$,

$$\begin{aligned}\|u(t_n) - u_h^n\|_L \le c'\,e^{c\frac{t_n}{\rho}}\Big(&\tau^3(t_n\rho)^{\frac12}\big(c_1^n(u) + \tilde c_1^n(f)\big)\\ &+ (t_n\beta)^{\frac12}h^{k+\frac12}c_2^n(u) + (\rho\beta)^{\frac12}h^{k+\frac12}c_3^n(u)\Big),\end{aligned} \tag{78.29}$$

with $c_1^n(u) := \|\partial_t^4 u\|_{C^0([0,t_n];L)}$, $c_2^n(u) := \sum_{q\in\{0:2\}}\rho^q|\partial_t^q u|_{C^0([0,t_n];H^{k+1})}$, $c_3^n(u) := |u(t_n)|_{H^{k+1}}$, *and* $\tilde c_1^n(f)$ *defined in Lemma* 78.21.

Proof. See Burman et al. [74, Thm. 4.1] for time-dependent Friedrichs' systems discretized using either stabilized H^1-conforming or discontinuous finite elements. See also Zhang and Shu [293, Thm. 5.1] for nonlinear scalar conservation laws, discontinuous finite elements, and the SSPRK(3,3) scheme. □

Exercises

Exercise 78.1 (Order conditions). (i) Consider the linear ODE system $\partial_t\mathsf{U} = \tilde{\mathcal{A}}\mathsf{U} + \tilde{\mathsf{F}}$. Let $p \ge 1$. Prove that

$$\mathsf{U}(t_n) = \sum_{r\in\{0:p\}}\frac{\tau^r}{r!}\tilde{\mathcal{A}}^r\mathsf{U}(t_{n-1}) + \tau\mathsf{G}_p(t_{n-1}) + \mathcal{O}(\tau^{p+1}), \tag{78.30}$$

with G_p defined in (78.13). (*Hint*: verify that $\partial_t^r\mathsf{U} = \tilde{\mathcal{A}}^r\mathsf{U} + \Phi_r(\tilde{\mathsf{F}})$ for all $r \ge 1$, with $\Phi_r(\tilde{\mathsf{F}}) := \sum_{q\in\{1:r\}}\tilde{\mathcal{A}}^{r-q}\partial_t^{q-1}\tilde{\mathsf{F}}$.) (ii) Let $\tilde{\mathsf{F}} \in C^\infty(\overline{J};\mathbb{C}^I)$. Consider the uncoupled ODE system $\partial_t\mathsf{U} = \tilde{\mathsf{F}}(t)$. Let $\mathsf{U}^{n-1} := \mathsf{U}(t_{n-1})$. Let U^n be given by the RK scheme. Show that a necessary and sufficient condition for $\mathsf{U}(t_n) - \mathsf{U}^n = \mathcal{O}(\tau^{p+1})$ is (78.10) with $r := 1$. (*Hint*: write a Taylor expansion of order $(p-1)$ of $\tilde{\mathsf{F}}(t_{n,j})$ for all $j \in \{1:s\}$.)

Exercise 78.2 (Condition (78.10)). (i) Show that if (78.9a) holds true, then $\sum_{j\in\{1:s\}}b_j(1-c_j)^m c_j^n = \frac{m!n!}{(m+n+1)!}$ for all $m,n \in \mathbb{N}$ s.t. $m+n \le p-1$. (*Hint*: recall that $(1+x)^m = \sum_{r\in\{0:m\}}\binom{m}{r}x^r$, $\frac{1}{n+l+1} = \int_0^1 x^{n+l}\,\mathrm{d}x$, and $\int_0^1(1-x)^m x^n\,\mathrm{d}x = \frac{m!n!}{(m+n+1)!}$.) (ii) Show that if (78.9a) and (78.9c) hold true, then $\sum_{i\in\{1:s\}}b_i(1-c_i)^{m-1}a_{ij} = \frac{b_j}{m}(1-c_j)^m$ for all $j \in \{1:s\}$ and all

$m \in \{1{:}\zeta\}$. (iii) Prove that (78.10) is met for $q := 1$ if (78.9a) and (78.9b) hold with $\eta := p-1$. (*Hint*: show that $\sum_{j_2,\dots,j_r\in\{1:s\}} a_{j_1j_2}\dots a_{j_{r-1}j_r} = \frac{1}{(r-1)!}c_{j_1}^{r-1}$ for all $r \in \{2{:}p\}$.) (iv) Prove that (78.10) is met for $q := 1$ if (78.9a) and (78.9c) hold with $\zeta := p-1$. (v) Show that (78.10) with $q := 1$ is met for all $r \in \{1{:}p\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $\eta+\zeta+1 = p$ (vi) Show that (78.10) is met for all $r \in \{1{:}p\}$ and all $q \in \{1{:}p-r+1\}$ if (78.9a) holds and (78.9b) and (78.9c) hold with $p \le \eta+\zeta+1$.

Exercise 78.3 (Explicit Euler). Revisit the proof of Lemma 78.12 by using the test function $w_h := u_h^n$ instead of $w_h := u_h^{n-1}$ and assuming that $\tau \le \min(\lambda_0\tau_2(h), \frac{1}{2}\frac{\rho}{1+\lambda_0\varpi^2})$ where $\varpi := \frac{h}{\beta}\sup_{v_h,w_h\in V_h}\frac{|a_h(v_h,w_h)|}{\|v_h\|_L\|w_h\|_L}$. (*Hint*: use that $a_h(u_h^{n-1},u_h^n) = a_h(u_h^n,u_h^n) + a_h(u_h^{n-1}-u_h^n,u_h^n)$.)

Exercise 78.4 (First-order viscosity). Let $(\cdot,\cdot)_V$ be a semidefinite Hermitian sesquilinear form in V and let $|\cdot|_V$ be the associated seminorm. Assume that $\Re\big((A(v),v)_L\big) \ge 0$ and $\|A(v)\|_L \le \beta\|v\|_L$ for all $v \in V$. Let $V_h \subset V$ and set $c_{\text{INV}}(h) := \max_{v_h\in V_h}\frac{|v_h|_V}{\|v_h\|_L}$. Given $u_h^0 \in V_h$, let $u_h^n \in V_h$ solve $\frac{1}{\tau}(u_h^n-u_h^{n-1},w_h)_L + (A(u^{n-1}),v_h)_L + \mu(u_h^{n-1},w_h)_V = 0$, for all $w_h \in V_h$ and all $n \in \mathcal{N}_\tau$, where $\mu \ge 0$ is an artificial viscosity parameter yet to be defined (μ can depend on h and τ). (i) Explain why this scheme can be more attractive than the implicit Euler method with $\mu := 0$. (ii) Prove that if $\tau(\beta+\mu c_{\text{INV}}(h))^2 \le 2\mu$, then $\|u_h^n\|_L \le \|u_h^0\|_L$ for all $n \in \mathcal{N}_\tau$. (iii) Prove that the above stability condition can be realized if and only if $2\beta\tau c_{\text{INV}}(h) \le 1$, and determine the admissible range for μ. *Note*: the constant $\beta\tau c_{\text{INV}}(h)$ is called Courant–Friedrichs–Levy (CFL) number.

Exercise 78.5 (Explicit Euler, mass lumping). Let $\beta \in \mathbb{R}$, $\beta \ne 0$. Consider the equation $\partial_t u + \beta\partial_x u = 0$ over $D := (0,1)$ with periodic boundary conditions. Use the same setting for the space discretization as in Exercise 77.1. (i) Write the linear system solved by the coordinate vector $(\mathsf{U}_1^n,\dots,\mathsf{U}_I^n)^\intercal$ by using the explicit Euler scheme and the Galerkin approximation with mass lumping. (*Hint*: use the convention $\mathsf{U}_I^n := \mathsf{U}_0^n$, $\mathsf{U}_{I+1}^n := \mathsf{U}_1^n$, $\mathsf{U}_{-1}^n := \mathsf{U}_{I-1}^n$.) (ii) Show that $\sum_{j\in\{1:I\}}(\mathsf{U}_j^n)^2 = \sum_{j\in\{1:I\}}(\mathsf{U}_j^{n-1})^2 + \lambda^2\sum_{j\in\{1:I\}}(\mathsf{U}_{j+1}^{n-1}-\mathsf{U}_{j-1}^{n-1})^2$ with $\lambda := \frac{\beta\tau}{2h}$. (iii) Let $a := (1-2\mathrm{i}\lambda\sin(\frac{k}{I}2\pi))$ where $k \in \mathbb{N}$ and $\frac{k}{I} \notin \mathbb{N}$, $\mathrm{i}^2 := -1$, and set $\mathsf{U}_j^0 := ae^{\mathrm{i}\frac{j}{I}2k\pi}$ for all $j \in \{1{:}I\}$. Compute U_j^n for all $n \in \mathcal{N}_\tau$ and comment on the result.

Exercise 78.6 (Error equation, RK2). (i) Verify that

$$u(t_n) = u(t_{n-1}) + \tau\partial_t u(t_{n-1}) + \frac{1}{2}\tau^2\partial_{tt}u(t_{n-1}) + \frac{1}{2}\tau\psi^{n-1},$$

with $\psi^{n-1} := \frac{1}{\tau}\int_{J_n}(t_n-t)^2\partial_{ttt}u(t)\mathrm{d}t$. (*Hint*: integrate by parts in time.) (ii) Prove (78.26). (*Hint*: use the fact that $(\partial_{tt}u(t_{n-1}),w_h)_L + (A(\partial_t u(t_{n-1})),w_h)_L = (\partial_t f^{n-1},w_h)_L$ for all $w_h \in V_h$.)

Exercise 78.7 (ERK schemes, $p=3$). Prove Lemma 78.21. (*Hint*: proceed as in the proof of Lemma 78.15, use that $\|A_h(w_h)\|_L \le c\frac{1}{\rho}\|w_h\|_{H^1}$ for all $w_h \in V_h$, and invoke the H^1-stability of $\mathcal{P}_{V_h}$ (see Proposition 22.21).)

Chapter 79

Scalar conservation equations

In Part XVI, composed of Chapters 79 to 83, we consider scalar conservation equations and hyperbolic systems. The first two chapters deal with the fundamental mathematical properties of such problems. The other three chapters deal with the finite element approximation, first using a low-order scheme and then extending it to higher order in time and in space. The present chapter gives a brief description of the theory of scalar conservation equations. We introduce the notions of weak and entropy solutions and state existence and uniqueness results. Even if the initial data is smooth, the solution of a generic scalar conservation equation may lose smoothness in finite time, and weak solutions are in general nonunique. Uniqueness is recovered by enforcing constraints that are called entropy conditions. We finish this chapter by exploring the structure of a one-dimensional Cauchy problem called Riemann problem where the initial data is composed of two constant states. Understanding the structure of the solution to the Riemann problem is important to understand the approximation techniques discussed in Chapter 81.

79.1 Weak and entropy solutions

In this section, we introduce the key notions of weak and entropy solutions.

79.1.1 The model problem

Let D be an open polyhedron in $\mathbb{R}^d$. Let $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ be a Lipschitz vector-valued function hereafter called *flux*, and let $u_0 \in L^\infty(D)$ be some initial data. We consider the scalar-valued conservation equation

$$\partial_t u + \nabla\cdot\boldsymbol{f}(u) = 0, \qquad u(\boldsymbol{x},0) = u_0(\boldsymbol{x}), \qquad (\boldsymbol{x},t) \in D\times\mathbb{R}_+, \tag{79.1}$$

where $\mathbb{R}_+ := [0,\infty)$. Problem (79.1) is called *Cauchy problem*. To simplify questions regarding boundary conditions, we assume that either periodic boundary conditions are enforced, or there is a compact subset $S \subsetneq D$ s.t. $u_{0|D\setminus S}$ is constant over each connected component of $D\setminus S$ (there is only one connected component if $d \geq 2$), and there exists some time $T > 0$ such that $u(\boldsymbol{x},t) = u_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial D$ and all $t \in [0,T]$. For more general boundary conditions, we refer the reader to Bardos et al. [23].

The PDE in (79.1) is called *conservation equation*. To better understand the origin of this terminology, let O be an open set in D. Then if u solves (79.1) we have $\partial_t \int_O u(\boldsymbol{x},t)\,\mathrm{d}x =$

$-\int_{\partial O} \boldsymbol{f}(u)\cdot\boldsymbol{n}\,\mathrm{d}s$. Calling $\int_O u(\boldsymbol{x},t)\,\mathrm{d}x$ the mass in O, this identity means that the rate of change of the mass is the opposite of the mass flux at the boundary of O. In particular, the mass in O is conserved over time if $\boldsymbol{f}(u)_{|\partial O}\cdot\boldsymbol{n}=0$.

Example 79.1 (Linear transport). The linear transport equation $\partial_t u+\boldsymbol{\beta}\cdot\nabla u=0$, where $\boldsymbol{\beta}\in\mathbb{R}^d$ is a constant vector field, can be recast into the form (79.1) by setting $\boldsymbol{f}(u):=\boldsymbol{\beta}u$. The solution to the Cauchy problem in $\mathbb{R}^d$ is $u(\boldsymbol{x},t)=u_0(\boldsymbol{x}-\boldsymbol{\beta}t)$, i.e., the graph of the solution at any time $t>0$ is the same as that of u_0, up to the translation $\boldsymbol{x}\mapsto\boldsymbol{x}-\boldsymbol{\beta}t$. □

Example 79.2 (Burgers' equation). In dimension one, the flux $\boldsymbol{f}(u):=\frac{1}{2}u^2\boldsymbol{e}_x$ gives Burgers' equation, where $\boldsymbol{e}_x$ is the unit vector giving the orientation of $\mathbb{R}$. The conservation equation is $\partial_t u+\frac{1}{2}\partial_x u^2=0$. □

Example 79.3 (Traffic flow equation). Setting $\boldsymbol{f}(u):=v_{\max}u(1-\frac{u}{u_{\max}})\boldsymbol{e}_x$ we obtain the traffic flow equation. This equation models automobile traffic on a one-lane road. Here, u is the number of cars per unit length (car density), $v_{\max}$ is the speed limit, and $u_{\max}$ is the maximum density observed in a traffic jam when all the cars are at rest bumper to bumper. The velocity of the cars is $\boldsymbol{v}(u):=\frac{1}{u}\boldsymbol{f}(u)=v_{\max}(1-\frac{u}{u_{\max}})\boldsymbol{e}_x$. When the density is small (u close to 0), the velocity is close to $v_{\max}\boldsymbol{e}_x$, which means that all the cars move along at the speed limit. When the density is close to maximum density (u close to $u_{\max}$), the velocity is close to $\boldsymbol{0}$, i.e., there is a traffic jam. □

Example 79.4 (Buckley–Leverett). The flow of a mixture of oil and water in a porous medium can be approximated by the Buckley–Leverett model. In this case, the dependent variable is the water saturation $u\in[0,1]$, and the flux is given by $\boldsymbol{f}(u):=\boldsymbol{\beta}\frac{u^2}{u^2+M(1-u)^2}$, where $\boldsymbol{\beta}$ is the total velocity, which we assume to be a constant field in $\mathbb{R}^d$, and $M>0$ is the ratio of the water viscosity to the oil viscosity. □

79.1.2 Short-time existence and loss of smoothness

In this section, we are concerned with smooth solutions to (79.1).

Definition 79.5 (Strong solution). *We say that u is a* strong solution *to* (79.1) *over the time interval $[0,T^*)$ for some $T^*>0$ if $u\in C^1(D\times[0,T^*))$ and u solves* (79.1) *for all $(\boldsymbol{x},t)\in D\times[0,T^*)$.*

Assuming that u is a strong solution, we can recast (79.1) as $\partial_t u+\boldsymbol{f}'(u)\cdot\nabla u=0$. In other words, looking for a strong solution to (79.1) is equivalent to solving a nonlinear transport equation with velocity $\boldsymbol{f}'(u)$. For Burgers' equation, we have $\boldsymbol{f}'(u)=u\boldsymbol{e}_x$, for the traffic flow equation, we have $\boldsymbol{f}'(u)=v_{\max}(1-\frac{2u}{u_{\max}})\boldsymbol{e}_x$, and for the Buckley–Leverett equation, we have $\boldsymbol{f}'(u)=\boldsymbol{\beta}\frac{2Mu(1-u)}{(u^2+M(1-u)^2)^2}$.

We are now going to show that an implicit representation of a strong solution to (79.1) can be obtained by the *method of characteristics* for short times in dimension one if $\boldsymbol{f}(u):=f(u)\boldsymbol{e}_x$ is of class C^2 and u_0 is of class C^1. It is not our goal here to give a detailed description of the method of characteristics. We are just going to outline the main idea which consists of considering the following ordinary differential equation:

$$\begin{cases}\partial_t\chi(s,t)=f'(u(\chi(s,t),t)), & t\ge 0,\\ \chi(s,0)=s,\end{cases}\tag{79.2}$$

where the parameter s spans $\mathbb{R}$ and u is assumed to be a smooth solution to (79.1). The curves $\{(x,t)\in\mathbb{R}\times\mathbb{R}_+\mid\chi(s,t)=x\}$ defined in the half plane $\mathbb{R}\times\mathbb{R}_+$ and parameterized by $s\in\mathbb{R}$ are called characteristics. After setting $\psi(s,t):=u(\chi(s,t),t)$, one observes that $\partial_t\psi(s,t)=0$, so that $u(\chi(s,t),t)=\psi(s,t)=\psi(s,0)=u(\chi(s,0),0)=u(s,0)=u_0(s)$, which in turn implies that

$\chi(s,t) = f'(u_0(s))t + s$. In conclusion, we have obtained an implicit representation of the strong solution to (79.1) in the form

$$u(\chi(s,t),t) = u_0(s),\ \forall s \in \mathbb{R} \quad \text{where } \chi(s,t) := f'(u_0(s))t + s. \tag{79.3}$$

For the linear transport equation where $f(u) = \beta u$, we have $\chi(s,t) = \beta t + s$, so that $s = \chi(s,t) - \beta t$. Hence, $u(\chi(s,t),t) = u_0(\chi(s,t) - \beta t)$ for all $s \in \mathbb{R}$. Since $\chi(\cdot,t) : \mathbb{R} \to \mathbb{R}$ is surjective (bijective actually), the above identity implies that $u(x,t) = u_0(x - \beta t)$ for all $x \in \mathbb{R}$. This argument shows that one can obtain an explicit representation of the strong solution if one can invert the map $\chi(\cdot,t) : \mathbb{R} \to \mathbb{R}$.

Let us suppose for a moment that there exists $T^* > 0$ such that $\chi(\cdot,t) : \mathbb{R} \to \mathbb{R}$ is invertible for all $t \in [0,T^*)$. Then we have

$$u(x,t) = u_0(\chi^{-1}(x,t)), \qquad \forall (x,t) \in \mathbb{R} \times [0,T^*). \tag{79.4}$$

The rest of the argument consists of proving that indeed there exists $T^* > 0$ such that $\chi(\cdot,t) : \mathbb{R} \to \mathbb{R}$ is invertible for all $t \in [0,T^*)$. Let $x \in \mathbb{R}$ and $t \geq 0$, and consider the equation

$$G(s,x,t) := x - f'(u_0(s))t - s = 0, \tag{79.5}$$

where s is the unknown. Using the implicit function theorem, we infer that the equation $G(s,x,t) = 0$ has a unique solution if $\partial_s G \neq 0$, i.e., if $f''(u_0(s))u_0'(s)t + 1 \neq 0$. If $f''(u_0(s))u_0'(s) \geq 0$ for all $s \in \mathbb{R}$, we set $T^* := \infty$. If there exists some s_0 s.t. $f''(u_0(s_0))u_0'(s_0) < 0$, we set $T^* := \inf_{s\in\mathbb{R}} \frac{-1}{\min(f''(u_0(s))u_0'(s),0)}$. Then, provided $T^* > 0$, we infer that for all $x \in \mathbb{R}$ and all $t \in [0,T^*)$, there is a unique $s \in \mathbb{R}$ such that $x = f'(u_0(s))t + s$, and we set $\chi^{-1}(x,t) := s$. Note that the implicit function theorem implies that χ^{-1} is of class C^1 w.r.t x and t. We refer the reader to Exercise 79.7 for other details. In conclusion, we have shown the following result.

Proposition 79.6 (Existence time for a strong solution). *Assume that f is of class C^2, u_0 is of class C^1, and $\inf_{s\in\mathbb{R}} \min(f''(u_0(s))u_0'(s),0) > -\infty$. Then* (79.1) *has a unique strong solution over the time interval $[0,T^*)$, where $T^* := \infty$ if $\inf_{s\in\mathbb{R}} f''(u_0(s))u_0'(s) \geq 0$ and otherwise we have*

$$T^* := \inf_{s\in\mathbb{R}} \frac{-1}{\min(f''(u_0(s))u_0'(s),0)} < \infty. \tag{79.6}$$

Example 79.7 (Burgers). Consider Burgers' equation, i.e., $f(u) := \frac{1}{2}u^2$. Then $f''(u_0(s))u_0'(s) = u_0'(s)$. Consider first the increasing function $u_0(s) := \tanh(s)$ as the initial condition. Then Proposition 79.6 leads to $T^* = \infty$, i.e., the strong solution exists at all times. But for the decreasing function $u_0(s) := -\tanh(s)$, we obtain $T^* = 1$, i.e., the strong solution exists in this case up to the time $T^* = 1$, and it turns out that no strong solution exists for longer times. □

The striking property here is that smoothness can be lost in finite time. To better understand this phenomenon, consider Burgers' equation with the initial data $u_0(x) := 1$ if $x < 0$, $u_0(x) := 1-x$ if $0 \leq x \leq 1$, and $u_0(x) := 0$ if $1 \leq x$, as shown in the bottom panel of Figure 79.1. Here, u_0 is not of class C^1, but it can be shown that the solution produced by the method of characteristics is still legitimate. Let us apply the method. For $s \leq 0$, we have $\chi(s,t) = t + s$, which gives $s = \chi^{-1}(x,t) = x - t$. Hence, $u(x,t) = u_0(s) = 1$ if $\chi^{-1}(x,t) \leq 0$, i.e., if $x \leq t$. For $0 \leq s \leq 1$, we have $\chi(s,t) = (1-s)t + s$, which gives $s = \chi^{-1}(x,t) = \frac{x-t}{1-t}$. Hence, $u(x,t) = 1 - s = \frac{1-x}{1-t}$ if $0 \leq \chi^{-1}(x,t) \leq 1$, i.e., if $t \leq x \leq 1$. For $1 \leq s$, we have $\chi(s,t) = s$, which gives $\chi^{-1}(x,t) = x$. Hence, $u(x,t) = u_0(s) = 0$ if $1 \leq \chi^{-1}(x,t)$, i.e., if $1 \leq x$. Note that $\chi^{-1}(x,t)$ is well defined for all $t \in [0,1)$, but the above solution is not well defined for $t = 1$. We have $u(x,1) = 1$ if $x \leq 1$,

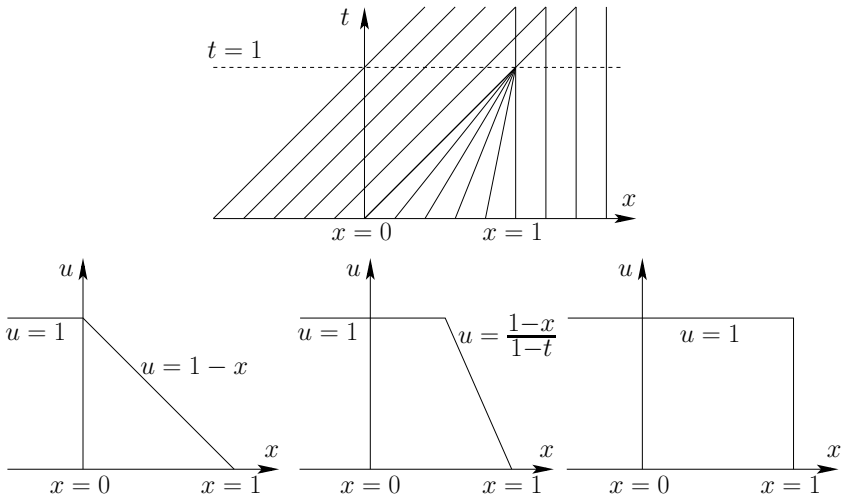

Figure 79.1: Top: characteristics for Burgers' equation. Bottom, from left to right: (i) solution at $t = 0$; (ii) solution at $t \in (0, 1)$; (iii) solution at $t = 1$.

$u(1, 1) = \frac{0}{0}$, and $u(x, 1) = 0$ if $1 \le x$, so that u is multivalued at $x = 1$. Actually, the solution is defined almost everywhere at $t = 1$, and we say that there is a shock at $x = 1$ at $t = 1$. Let us compute T^*. We have $f''(u_0(s))u_0'(s) = u_0'(s)$, $u_0'(s) = 0$ if $s < 0$ and $s > 1$, and $u_0'(s) = -1 < 0$ if $0 < s < 1$. This computation shows that $T^* = 1$, which indeed is the time when the solution given by the method of characteristics produces a shock. The characteristics for this problem are shown in the top panel of Figure 79.1. We observe that the solution is multivalued at $x = 1$ when $t = 1$, i.e., many characteristics cross at this point. This feature is generic: for every flux, the solution given by the method of characteristics ceases to make sense once some characteristics cross.

79.1.3 Weak solutions

In order to make sense of solutions to (79.1) that are not of class C^1, because either the initial data is not of class C^1 or smoothness is lost at some time T^*, we now introduce the notion of weak solutions. A weak formulation of (79.1) is obtained by testing the equation with smooth test functions that are compactly supported in $D\times\mathbb{R}_+$, say $\phi \in C_0^1(D\times\mathbb{R}_+)$, integrating over the space-time domain $D\times\mathbb{R}_+$, and integrating by parts as follows:

$$\int_0^\infty \int_D (u\partial_t\phi + \boldsymbol{f}(u)\cdot\nabla\phi)\,\mathrm{d}x\mathrm{d}t + \int_D \phi(\boldsymbol{x}, 0)u_0(\boldsymbol{x})\,\mathrm{d}x = 0. \tag{79.7}$$

Since $\mathbb{R}_+ := [0, \infty)$, $\phi(0, \cdot)$ can be nonzero over a compact subset of the line $D\times\{t{=}0\}$ (see Definition 1.31 for the notion of support). Moreover, the space $L^\infty_{\mathrm{loc}}(D\times\mathbb{R}_+)$ is by definition composed of functions that are bounded on each compact subset of $D\times\mathbb{R}_+$.

Definition 79.8 (Weak solution). *We say that $u \in L^\infty_{\mathrm{loc}}(D\times\mathbb{R}_+)$ is a* weak solution *to* (79.1) *if u satisfies* (79.7) *for all $\phi \in C_0^1(D\times\mathbb{R}_+)$.*

If u is smooth and is a weak solution to (79.7), then restricting the test functions in (79.7) to $C_0^\infty(D\times(0, \infty))$ shows that u solves $\partial_t u + \nabla\cdot\boldsymbol{f}(u) = 0$.

Example 79.9 (Linear transport). Assuming that $D := \mathbb{R}^d$ and $u_0 \in L^\infty_{\mathrm{loc}}(D)$, let us show that $u(\boldsymbol{x}, t) = u_0(x - \boldsymbol{\beta}t)$ is indeed a weak solution to the linear transport equation $\partial_t u + \nabla\cdot(\boldsymbol{\beta}u) = 0$,

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is a constant vector field. Let us denote $\mathfrak{T} := \int_0^\infty \int_D (u\partial_t\phi(\boldsymbol{x},t) + u\boldsymbol{\beta}\cdot\nabla\phi(\boldsymbol{x},t))\,\mathrm{d}x\mathrm{d}t$, and let us make the change of variable $\boldsymbol{x}' = \boldsymbol{x} - \boldsymbol{\beta}t$. We obtain

$$\mathfrak{T} = \int_0^\infty \int_D \big(u_0(\boldsymbol{x}')\partial_t\phi(\boldsymbol{x}'+\boldsymbol{\beta}t,t) + u_0(\boldsymbol{x}')\boldsymbol{\beta}\cdot\nabla\phi(\boldsymbol{x}'+\boldsymbol{\beta}t,t)\big)\,\mathrm{d}x'\mathrm{d}t.$$

For all $\boldsymbol{x}' \in D$, let us set $\psi(\boldsymbol{x}',t) := \phi(\boldsymbol{x}'+\boldsymbol{\beta}t,t)$. Then $\partial_t\psi(\boldsymbol{x}',t) = \boldsymbol{\beta}\cdot\nabla\phi(\boldsymbol{x}'+\boldsymbol{\beta}t,t) + \partial_t\phi(\boldsymbol{x}'+\boldsymbol{\beta}t,t)$. Applying Fubini's theorem gives $\mathfrak{T} = \int_D u_0(\boldsymbol{x}')\int_0^\infty \partial_t\psi(\boldsymbol{x}',t)\mathrm{d}t\,\mathrm{d}x' = \int_D -u_0(\boldsymbol{x})\psi(\boldsymbol{x},0)\,\mathrm{d}x = \int_D -u_0(\boldsymbol{x})\phi(\boldsymbol{x},0)\,\mathrm{d}x$. In conclusion, the identity (79.7) holds true. □

In general, there are infinitely many weak solutions to (79.1). Consider for instance Burgers' equation in dimension one with $u_0(x) := H(x)$, where H is the Heaviside function (i.e., $H(x) := 1$ if $x \geq 0$ and $H(x) := 0$ if $x < 0$). Let us verify that the following two functions:

$$u_1(x,t) := H(x - \tfrac{1}{2}t) \quad \text{and} \quad u_2(x,t) := \begin{cases} 0 & \text{if } x < 0, \\ \frac{x}{t} & \text{if } 0 < x < \mathrm{t}, \\ 1 & \text{if } x > t, \end{cases} \tag{79.8}$$

are weak solutions, that is, let us show that (79.7) holds true with $D := \mathbb{R}$ in both cases for every test function $\phi \in C_0^1(\mathbb{R}\times\mathbb{R}_+)$. Let us denote by $\mathfrak{T}_1$ the double integral on the left-hand side of (79.7) with $u := u_1$. Using Fubini's theorem for the double integral involving $\partial_t\phi$, we obtain

$$\begin{aligned}
\mathfrak{T}_1 &= \int_{\mathbb{R}}\int_0^\infty H(x-\tfrac{1}{2}t)\partial_t\phi\mathrm{d}t\,\mathrm{d}x + \int_0^\infty\int_{\mathbb{R}} \frac{1}{2}H^2(x-\tfrac{1}{2}t)\partial_x\phi\,\mathrm{d}x\mathrm{d}t \\
&= \int_0^\infty \int_0^{2x} \partial_t\phi\mathrm{d}t\,\mathrm{d}x + \int_0^\infty\int_{\frac{t}{2}}^\infty \frac{1}{2}\partial_x\phi\,\mathrm{d}x\mathrm{d}t \\
&= \int_0^\infty (\phi(x,2x) - \phi(x,0))\,\mathrm{d}x - \frac{1}{2}\int_0^\infty \phi(\frac{t}{2},t)\mathrm{d}t = -\mathfrak{T}_0,
\end{aligned}$$

with $\mathfrak{T}_0 := \int_0^\infty \phi(x,0)\,\mathrm{d}x = \int_{\mathbb{R}} u_0(x)\phi(x,0)\,\mathrm{d}x$ since $u_0(x) = 0$ if $x < 0$. Let us denote by $\mathfrak{T}_2$ the double integral on the left-hand side of (79.7) with $u := u_2$. Then $\mathfrak{T}_2 := \mathfrak{T}_{2,1} + \mathfrak{T}_{2,2}$ with

$$\begin{aligned}
\mathfrak{T}_{2,1} &:= \int_0^\infty \Big(\int_0^x \partial_t\phi\mathrm{d}t + \int_x^\infty \frac{x}{t}\partial_t\phi\mathrm{d}t\Big)\,\mathrm{d}x \\
&= \int_0^\infty -\phi(x,0)\,\mathrm{d}x + \int_0^\infty\int_x^\infty \frac{x}{t^2}\phi\mathrm{d}t\,\mathrm{d}x,
\end{aligned}$$

where we used Fubini's theorem and integrated by parts in time, and

$$\mathfrak{T}_{2,2} := \int_0^\infty \Big(\int_0^t \frac{1}{2}\frac{x^2}{t^2}\partial_x\phi\,\mathrm{d}x + \int_t^\infty \frac{1}{2}\partial_x\phi\,\mathrm{d}x\Big)\mathrm{d}t = \int_0^\infty\int_0^t -\frac{x}{t^2}\phi\,\mathrm{d}x\mathrm{d}t,$$

where we integrated by parts in space. Invoking once again Fubini's theorem and observing that $\{x \in \mathbb{R}_+,\, t \geq x\} = \{t \in \mathbb{R}_+,\, x \in (0,t)\}$ leads to $\mathfrak{T}_2 = \mathfrak{T}_{2,1} + \mathfrak{T}_{2,2} = -\mathfrak{T}_0$. The reader is referred to the Exercises 79.2 and 79.3 for more details on the uniqueness question.

79.1.4 Existence and uniqueness

The nonuniqueness problem can be solved by invoking additional considerations on viscous dissipation. We say that u is a physically relevant solution to (79.1) if it is a weak solution and if it is

the limit in some appropriate topology of the unique solution to the following perturbed problem as $\epsilon \to 0$:

$$\partial_t u_\epsilon + \nabla\cdot\boldsymbol{f}(u_\epsilon) - \epsilon\Delta u_\epsilon = 0, \qquad u_\epsilon(\boldsymbol{x},0) = u_0(\boldsymbol{x}), \quad (\boldsymbol{x},t) \in D\times\mathbb{R}_+. \tag{79.9}$$

We say that u_ϵ is the *viscous regularization* of u or the viscous approximation to (79.1). The limiting process has been studied in detail in Oleĭnik [232, 233], Kružkov [206], where it is proved that requesting that a weak solution to (79.7) be such that $\lim_{\epsilon\to 0}\|u_\epsilon - u\|_{L^1(D\times(0,T);\mathbb{R})} = 0$ is equivalent to requiring that u satisfy the additional entropy inequalities $\partial_t\eta(u) + \nabla\cdot\boldsymbol{q}(u) \le 0$ (in the distribution sense) for any convex function $\eta \in \mathrm{Lip}(\mathbb{R};\mathbb{R})$ with associated flux $\boldsymbol{q} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ s.t. $q_l(u) := \int_0^u \eta'(v) f_l'(v)\,\mathrm{d}v$ for all $l \in \{1\!:\!d\}$. The functions η and $\boldsymbol{q}$ are called *entropy* and *entropy flux*.

Theorem 79.10 (Entropy solution). *Let $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$ and $u_0 \in L^\infty(D)$. Let the assumptions on the boundary conditions stated in §79.1.1 hold true. There is a unique entropy solution to (79.1), i.e., there is a unique function $u \in L^\infty_{\mathrm{loc}}(D\times\mathbb{R}_+)$ that is a weak solution and that satisfies the following* entropy inequalities*:*

$$-\int_0^\infty\int_D (\eta(u)\partial_t\phi + \boldsymbol{q}(u)\cdot\nabla\phi)\,\mathrm{d}x\mathrm{d}t - \int_D \phi(x,0)\eta(u_0)\,\mathrm{d}x \le 0, \tag{79.10}$$

for all the entropy pairs $(\eta,\boldsymbol{q})$ and all $\phi \in C_0^1(D\times\mathbb{R}_+;\mathbb{R}_+)$ (note that ϕ here takes nonnegative values). In other words, we have $\partial_t\eta(u) + \nabla\cdot\boldsymbol{q}(u) \le 0$ in the distribution sense in $D\times(0,\infty)$.

Theorem 79.11 (Maximum principle). *Let us set $u_{\min} := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} u_0(\boldsymbol{x})$ and $u_{\max} := \operatorname{ess\,sup}_{\boldsymbol{x}\in D} u_0(\boldsymbol{x})$. The entropy solution satisfies the following maximum principle:*

$$u(\boldsymbol{x},t) \in [u_{\min}, u_{\max}], \qquad \text{for a.e. } (\boldsymbol{x},t) \in D\times\mathbb{R}_+. \tag{79.11}$$

Remark 79.12 (Kružkov entropies). It can be shown that Theorem 79.10 holds true if the inequality (79.10) is satisfied only for the Kružkov entropies $\eta_k(u) := |u-k|$, with flux $\boldsymbol{q}_k(u) := \mathrm{sign}(u-k)(\boldsymbol{f}(u) - \boldsymbol{f}(k))$ for all $k \in [u_{\min}, u_{\max}]$; see Exercise 79.1. □

Remark 79.13 (Strong solution). Strong solutions are also weak solutions and they satisfy all the entropy inequalities. This follows from the definition of the entropy flux and the chain rule. See also Exercise 79.6. □

79.2 Riemann problem

In this section, we introduce the notion of Riemann problem and give a brief overview of the construction of its solution. Understanding the structure of the solution to the Riemann problem is important to understand the technique presented in Chapter 81 to approximate the Cauchy problem (79.1). In the entire section, we assume that $\boldsymbol{f}$ is at least Lipschitz, i.e., $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$.

79.2.1 One-dimensional Riemann problem

The Riemann problem is a particular instance of the Cauchy problem (79.1), where the space is one-dimensional and the initial data consists of two constant states. More precisely, setting $\boldsymbol{f}(v) := f(v)\boldsymbol{e}_x$, the *Riemann problem* consists of solving the following Cauchy problem:

$$\partial_t u + \partial_x f(u) = 0, \qquad u(x,0) := \begin{cases} u_L & \text{if } x < 0,\\ u_R & \text{if } x > 0, \end{cases} \tag{79.12}$$

with $u_L, u_R \in \mathbb{R}$. Since the solution to (79.12) is trivial if $u_L = u_R$, we focus on the case $u_L \neq u_R$. The key idea is that the solution to (79.12) is self-similar, i.e., it only depends on the ratio $\frac{x}{t}$. In other words, there is a function $w : \mathbb{R} \to \mathbb{R}$ such that $u(x,t) := w(\frac{x}{t})$. The motivation for looking for a solution of this form is the observation that if $u(x,t)$ solves (79.12), then $u_\lambda(x,t) = u(\lambda x, \lambda t)$ also solves (79.12) for all $\lambda > 0$. After setting $\xi := \frac{x}{t}$ and inserting the ansatz $u(x,t) = w(\xi)$ into (79.12), one obtains $-\frac{x}{t^2}w'(\xi) + \frac{1}{t}f'(w(\xi))w'(\xi) = 0$, so that $u(x,t) = w(x/t)$ solves (79.12) iff w satisfies the identity

$$\xi = f'(w(\xi)). \tag{79.13}$$

Solving this nonlinear equation requires that we investigate the monotonicity properties of f'.

79.2.2 Convex or concave flux

If f' is strictly monotone, then $f' : \mathbb{R} \to \mathbb{R}$ is invertible and the solution to $\xi = f'(w(\xi))$ is $w(\xi) = (f')^{-1}(\xi)$. Let us now make sense of this argument.

Let us assume that $u_L < u_R$. Assume that f is of class C^2 and strictly convex in the interval $[u_L, u_R]$. Then both $f' : [u_L, u_R] \to \mathbb{R}$ and $(f')^{-1} : [f'(u_L), f'(u_R)] \to \mathbb{R}$ are monotonically increasing. Since for every $t \geq 0$ the viscous solution to (79.12) is monotone in x (see Holden and Risebro [184, §2.1]), we connect u_L to u_R with a monotone increasing profile by setting

$$u(x,t) := \begin{cases} u_L & \text{if } \frac{x}{t} \leq f'(u_L), \\ (f')^{-1}(\frac{x}{t}) & \text{if } f'(u_L) < \frac{x}{t} \leq f'(u_R), \\ u_R & \text{if } f'(u_R) < \frac{x}{t}. \end{cases} \tag{79.14}$$

It can be proved that this is indeed the entropy solution to (79.12) (see [184, §2.2]). This solution is called *expansion wave*. The above argument does not make sense if f is strictly concave, since in this case $f'(u_L) > f'(u_R)$. It can then be shown that the correct solution is a discontinuity moving with the velocity $s := \frac{f(u_L)-f(u_R)}{u_L-u_R}$, i.e.,

$$u(x,t) := \begin{cases} u_L & \text{if } \frac{x}{t} \leq s, \\ u_R & \text{if } s < \frac{x}{t}, \end{cases} \qquad s := \frac{f(u_L)-f(u_R)}{u_L-u_R}. \tag{79.15}$$

This solution is called *shock wave* or simply *shock*. Graphical representations of the expansion wave and the shock wave are shown in Figure 79.2.

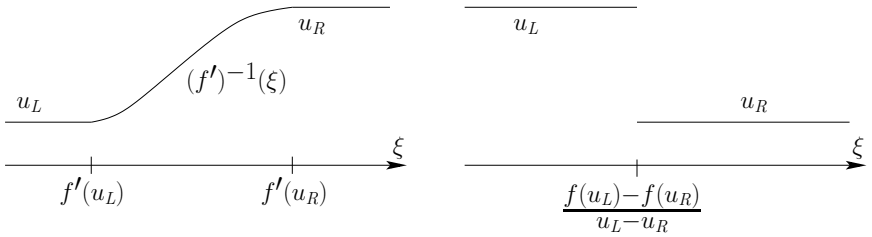

Figure 79.2: Solution to the Riemann problem $u(x,t) = w(\xi)$ when f is strictly convex. From left to right: (i) expansion wave; (ii) shock.

Recalling that for the time being we have assumed that $u_L < u_R$, the expansion wave (79.14) and the shock wave (79.15) can be recast into a single formalism by introducing the lower convex envelope of f over $[u_L, u_R]$:

$$\underline{f}(v) := \sup\{g(v) \mid g(z) \leq f(z), \forall z \in [u_L, u_R],\ g \text{ convex}\}. \tag{79.16}$$

To visualize the graph of $\underset{\sim}{f}$, think of a rubber band in $\mathbb{R}^2$ fixed at $(u_L, f(u_L))$ and $(u_R, f(u_R))$ and passing underneath the graph of f. This definition implies that $\underset{\sim}{f}(v) = f(u_L)\frac{v-u_R}{u_L-u_R} + f(u_R)\frac{v-u_L}{u_R-u_L}$ if f is concave and $\underset{\sim}{f}(v) = f(v)$ if f is convex. The two expressions (79.14) and (79.15) can be recast into a single formalism as follows:

$$u(x,t) := \begin{cases} u_L & \text{if } \frac{x}{t} \le \underset{\sim}{f}'(u_L), \\ (\underset{\sim}{f}')^{-1}(\frac{x}{t}) & \text{if } \underset{\sim}{f}'(u_L) < \frac{x}{t} \le \underset{\sim}{f}'(u_R), \\ u_R & \text{if } \underset{\sim}{f}'(u_R) < \frac{x}{t}. \end{cases} \tag{79.17}$$

Note that if f is concave, $\underset{\sim}{f}'(u_L) = \underset{\sim}{f}'(u_R) = s$ and the measure of the set $(\underset{\sim}{f}'(u_L), \underset{\sim}{f}'(u_R)]$ is zero, i.e., one does not have to bother to define $(\underset{\sim}{f}')^{-1}(s)$.

One treats the situation $u_L > u_R$ similarly by invoking the change of variable $x \to -x$ and $f \to -f$, but in this case the lower convex envelope of $-f$ is the upper concave envelope of f over $[u_R, u_L]$ defined by

$$\overline{f}(v) := \inf\{g(v) \mid f(z) \le g(z), \forall z \in [u_R, u_L],\ g \text{ concave}\}. \tag{79.18}$$

To visualize the graph of $\overline{f}$, think of a rubber band in $\mathbb{R}^2$ fixed at $(u_L, f(u_L))$ and $(u_R, f(u_R))$ and passing above the graph of f. The solution is defined by setting $u(x,t) := u_L$ if $\frac{x}{t} \le \overline{f}'(u_L)$, $u(x,t) := (\overline{f}')^{-1}(\frac{x}{t})$ if $\overline{f}'(u_L) < \frac{x}{t} \le \overline{f}'(u_R)$, and $u(x,t) := u_R$ if $\overline{f}'(u_R) < \frac{x}{t}$.

Remark 79.14 (Rankine–Hugoniot). When the solution to (79.12) is a shock wave, the identity $s = \frac{f(u_L)-f(u_R)}{u_L-u_R}$ is called *Rankine–Hugoniot condition* and s is called *shock speed*. □

79.2.3 General case

It turns out that the above argumentation can be generalized to any Lipschitz flux with finitely many inflection points.

Theorem 79.15 (Riemann solution). *Assume that the interval $[u_L, u_R]$ can be divided into finitely many subintervals where f has a continuous and bounded second derivative, and where f is either strictly convex or strictly concave. The entropy solution to* (79.12) *is given by*

$$u(x,t) := \begin{cases} u_L & \text{if } \frac{x}{t} \le \underset{\sim}{f}'(u_L), \\ (\underset{\sim}{f}')^{-1}(\frac{x}{t}) & \text{if } \underset{\sim}{f}'(u_L) < \frac{x}{t} \le \underset{\sim}{f}'(u_R), \\ u_R & \text{if } \underset{\sim}{f}'(u_R) < \frac{x}{t}, \end{cases} \tag{79.19}$$

if $u_L < u_R$, and $\underset{\sim}{f}$ must be replaced by $\overline{f}$ in (79.19) *if $u_L > u_R$.*

Proof. See Dafermos [96, Lem. 3.1] for the construction of the solution to the Riemann problem assuming that the flux is piecewise linear. See Holden and Risebro [184, §2.2] for a detailed proof. We refer to Osher [234, Thm. 1] for another interesting representation of the solution. □

79.2.4 Riemann cone and averages

Let $\lambda_L(u_L, u_R)$ and $\lambda_R(u_L, u_R)$ be the two quantities defined by

$$\lambda_L(u_L,u_R) := \begin{cases} \underset{\sim}{f}'(u_L) & \text{if } u_L < u_R, \\ \overline{f}'(u_L) & \text{if } u_L > u_R, \end{cases} \quad \lambda_R(u_L,u_R) := \begin{cases} \underset{\sim}{f}'(u_R) & \text{if } u_L < u_R, \\ \overline{f}'(u_R) & \text{if } u_L > u_R. \end{cases}$$

We will refer to $\lambda_L(u_L, u_R)$ and $\lambda_R(u_L, u_R)$ as the left and right extreme wave speeds, respectively. An important piece of information that we learn from Theorem 79.15 is that the solution to the Riemann problem is nontrivial in the space-time cone

$$C(u_L, u_R) := \left\{(x,t) \in \mathbb{R}\times\mathbb{R}_+ \mid \lambda_L(u_L, u_R) < \frac{x}{t} \le \lambda_R(u_L, u_R)\right\}. \tag{79.20}$$

It is equal to u_L on the left of $C(u_L, u_R)$ and equal to u_R on the right of $C(u_L, u_R)$. The cone $C(u_L, u_R)$ is often termed *Riemann fan* in the literature. A schematic representation of the Riemann fan is shown in Figure 79.3.

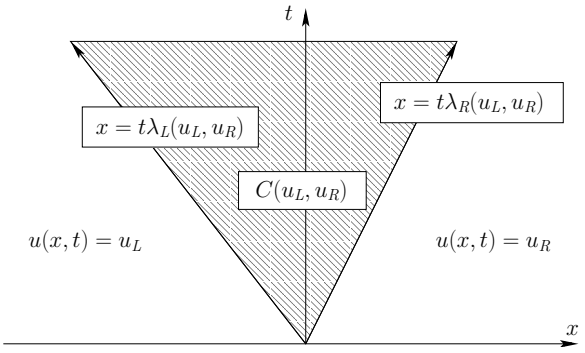

Figure 79.3: Riemann fan $C(u_L, u_R)$.

Definition 79.16 (Maximum wave speed). *We call* maximum wave speed *in the Riemann problem the number* $\max(|\lambda_L(u_L, u_R)|, |\lambda_R(u_L, u_R)|)$. *Any real number* $\lambda_{\max}(u_L, u_R)$ *satisfying the inequality*

$$\lambda_{\max}(u_L, u_R) \ge \max(|\lambda_L(u_L, u_R)|, |\lambda_R(u_L, u_R)|) \tag{79.21}$$

is called upper bound on the maximum wave speed.

The motivation for the above definition is that it is often easier to estimate an upper bound on $\max(|\lambda_L(u_L, u_R)|, |\lambda_R(u_L, u_R)|)$ than computing this quantity. For instance, if $f(v) := \cos(v)$, one can take $\lambda_{\max}(u_L, u_R) := 1$, but computing $\max(|\lambda_L(u_L, u_R)|, |\lambda_R(u_L, u_R)|)$ may not be simple.

Example 79.17 (Convex flux). Assume that f is convex. Then the quantity $\lambda_{\max}(u_L, u_R)$ defined by $\lambda_{\max}(u_L, u_R) := \left|\frac{f(u_L)-f(u_R)}{u_L-u_R}\right|$ if $u_L > u_R$, and $\lambda_{\max}(u_L, u_R) := \max(|f'(u_L)|, |f'(u_R)|)$ otherwise, satisfies (79.21). □

Lemma 79.18 (Riemann average). *Let* u *be the entropy solution to* (79.12), (η, q) *be an entropy pair, and define the* Riemann average *as* $\overline{u}(t, u_L, u_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} u(x,t)\,\mathrm{d}x$. *Let* $\lambda_{\max}(u_L, u_R)$ *be any upper bound on the maximum wave speed. Then for all* $t \in [0, \frac{1}{2\lambda_{\max}(u_L, u_R)}]$,

$$\overline{u}(t, u_L, u_R) = \frac{1}{2}(u_L + u_R) - t\big(f(u_R) - f(u_L)\big), \tag{79.22a}$$

$$\eta(\overline{u}(t, u_L, u_R)) \le \frac{1}{2}(\eta(u_L) + \eta(u_R)) - t\big(q(u_R) - q(u_L)\big). \tag{79.22b}$$

Proof. Let us integrate the equation $\partial_t u + \partial_x f(u) = 0$ over the domain $(-\frac{1}{2}, \frac{1}{2})\times(0,t)$, where u is the solution defined in Theorem 79.15. We obtain

$$\overline{u}(t, u_L, u_R) - \frac{1}{2}(u_L + u_R) + \int_0^t f(u(\tfrac{1}{2}, \tau))\,\mathrm{d}\tau - \int_0^t f(u(-\tfrac{1}{2}, \tau))\,\mathrm{d}\tau = 0.$$

Since $t \le \frac{1}{2\lambda_{\max}(u_L,u_R)}$, we have $u(-\frac{1}{2},\tau) = u_L$ because $-\frac{1}{2\tau} \le -\frac{1}{2t} \le -\lambda_{\max}(u_L,u_R) \le \lambda_L(u_L,u_R)$ for all $\tau \in (0,t]$, and we have $u(\frac{1}{2},\tau) = u_R$ because $\frac{1}{2\tau} \ge \lambda_R(u_L,u_R)$ for all $\tau \in (0,t]$. This proves (79.22a). The same argument applied to the inequality $\partial_t \eta(u) + \partial_x q(u) \le 0$ gives

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(u(x,t))\,\mathrm{d}x \le \tfrac{1}{2}(\eta(u_L)+\eta(u_R)) - t(q(u_R)-q(u_L)).$$

Jensen's inequality $\eta\big(\int_{-\frac{1}{2}}^{\frac{1}{2}} u(x,t)\,\mathrm{d}x\big) \le \int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(u(x,t))\,\mathrm{d}x$ gives (79.22b). □

Remark 79.19 (Invariant set/maximum principle). The maximum principle from Theorem 79.11 implies that u is in the convex hull of (u_L,u_R). This in turn implies that this is also the case of $\overline{u}(t,u_L,u_R)$. The identity (79.22a) says that for all $t \in [0, \frac{1}{2\lambda_{\max}(u_L,u_R)}]$,

$$\tfrac{1}{2}(u_L+u_R) - t\big(f(u_R)-f(u_L)\big) \in \operatorname{conv}(u_L,u_R).$$

This property is essential, and it will be used repeatedly in Chapter 81. □

79.2.5 Multidimensional flux

In Chapter 81, where we introduce an approximation technique for (79.1) with a multidimensional flux $\boldsymbol{f} \in \mathrm{Lip}(\mathbb{R};\mathbb{R}^d)$, we will consider the following one-dimensional Riemann problem: Find u such that

$$\partial_t u + \partial_x(\boldsymbol{f}(u)\cdot\boldsymbol{n}) = 0, \qquad u(x,0) := \begin{cases} u_L & \text{if } x<0,\\ u_R & \text{if } x>0, \end{cases} \tag{79.23}$$

where $u_L,u_R \in \mathbb{R}$ and $\boldsymbol{n}$ is an arbitrary unit vector in $\mathbb{R}^d$. The theory developed above can be directly applied to this case by setting $f(u) := \boldsymbol{f}(u)\cdot\boldsymbol{n}$. In this case, we denote by $\lambda_{\max}(\boldsymbol{n},u_L,u_R)$ any upper bound on the maximum wave speed in the Riemann problem (79.23).

Lemma 79.20 (Entropy pair). *Let $(\eta,\boldsymbol{q})$ be an entropy pair for* (79.1)*. Then $(\eta,\boldsymbol{q}\cdot\boldsymbol{n})$ is an entropy pair for the Riemann problem* (79.23)*.*

Proof. The identity $q_l(u) = \int_0^u \eta'(v) f_l'(v)\,\mathrm{d}v$, for all $l \in \{1{:}d\}$, implies that

$$\boldsymbol{q}(u)\cdot\boldsymbol{n} = \int_0^u \eta'(v)(\boldsymbol{f}\cdot\boldsymbol{n})'(v)\,\mathrm{d}v,$$

which proves the result. □

Exercises

Exercise 79.1 (Kružkov entropy pairs). For all $k \in \mathbb{R}$, consider the entropy $\eta(v,k) := |v-k|$. Compute the entropy flux associated with this entropy, $\boldsymbol{q}(v)$, with the normalization $\boldsymbol{q}(k) := \boldsymbol{0}$.

Exercise 79.2 (Entropy solution). Consider Burgers' equation with $D := \mathbb{R}$ and $u_0(x) := 0$. (i) What should be the entropy solution to this problem? (ii) Let H be the Heaviside function. Let $a \in \mathbb{R}$ and consider $u(x,t) := 2aH(x) - aH(x-\frac{at}{2}) - aH(x+\frac{at}{2})$. Draw the graph of $u(\cdot,t)$ at some time $t>0$. (iii) Show that u is a weak solution for all $a \in \mathbb{R}$. (iv) Verify that u is not the entropy solution. (*Hint*: consider the entropy $\eta(v) := |v|$.)

Exercise 79.3 (Entropy solution). Consider Burgers' equation with $D := \mathbb{R}$ and $u_0(x) := H(x)$, where H is the Heaviside function. (i) Verify that $u_1(x,t) := H(x - \frac{1}{2}t)$ and $u_2(x,t) := 0$ if $x < 0$, $u_2(x,t) := \frac{x}{t}$, if $0 < x < t$, $u_2(x,t) := 1$ if $x > t$, are both weak solutions. (ii) Verify that u_1 does not satisfy the entropy inequalities, whereas u_2 does.

Exercise 79.4 (Average speed). Let f be a scalar Lipschitz flux. Consider the Riemann problem $\partial_t u + \partial_x f(u) = 0$, with initial data (u_L, u_R), $u_L \neq u_R$. Let $\lambda_{\max}(u_L, u_R)$ be a maximum wave speed in this problem. Let $s := (f(u_L) - f(u_R))/(u_L - u_R)$ be the average speed. Assume that the interval $[u_L, u_R]$ can be divided into finitely many intervals where f has a continuous and bounded second derivative and f is either strictly convex or strictly concave. Prove that $|\lambda_{\max}(u_L, u_R)| \geq |s|$.

Exercise 79.5 (Maximum speed). Compute $\lambda_{\max}(u_L, u_R)$ for the two cases $(u_L, u_R) := (1, 2)$ and $(u_L, u_R) := (2, 1)$ with the following fluxes: (i) $f(v) := \frac{1}{2}v^2$; (ii) $f(v) := 8(v - \frac{1}{2})^3$; (iii) $f(v) := -(v-1)(2v-3)$ if $v \leq \frac{3}{2}$ and $f(v) := \frac{1}{4}(3 - 2v)$ if $\frac{3}{2} \leq v$.

Exercise 79.6 (Strong solutions). The goal is to justify Remark 79.13. (i) Show that if u is a weak solution and $u \in C^1(D \times [0, T^*))$, then u is a strong solution in $D \times [0, T^*)$. (ii) Show that if u is a strong solution, then u is also a weak solution. (iii) Let u be a strong solution to (79.1) and let $(\eta, \boldsymbol{q})$ an entropy pair with η of class C^2. Show that (79.10) holds true.

Exercise 79.7 (Method of characteristics). Let $D := \mathbb{R}$, $\boldsymbol{f} := f\boldsymbol{e}_x$, and assume that f is of class C^2 and u_0 is of class C^1. Recall that there exists $T^* > 0$ and a unique $s(x,t)$ solving $x = f'(u_0(s))t + s$ for all x and all $t \in [0, T^*)$. (i) Show that $u(x,t) := u_0(s(x,t))$ solves (79.1) for all $t \in [0, T^*)$. (ii) Let $s_0 \in \mathbb{R}$. Show that $u(x,t)$ is constant along the straight segment $\{x = f'(u_0(s_0))t + s_0 \mid t \in [0, T^*]\}$. (iii) Show that the solution found in Step (i) is the entropy solution.

Exercise 79.8 (Shock interacting with an expansion wave). Consider Burgers' equation with the initial condition $u_0(x) := -1$ if $x \in (-1, 0)$ and $u_0(x) := 0$ otherwise. (i) Derive the weak entropy solution up to the time $t = 2$. (ii) After the time $t = 2$, the shock originating from $x = -1$ starts interacting with the expansion wave originating from $x = 0$, leading to a shock with a nonlinear trajectory. Derive the weak entropy solution for the times $t \geq 2$. (*Hint*: use the Rankine–Hugoniot condition.) (iii) Verify that "mass" conservation is satisfied, i.e., $\int_{\mathbb{R}} u(x,t)\,\mathrm{d}x = \int_{\mathbb{R}} u_0(x)\,\mathrm{d}x = -1$ for all $t \geq 0$.

Chapter 80

Hyperbolic systems

The objective of this chapter is to introduce the concept of hyperbolic systems and to generalize the notions introduced in Chapter 79 to this class of equations. The novelty here is that the notion of maximum principle is no longer valid and is replaced by the concept of invariant sets. The material from this chapter is inspired from Bouchut [40, Chap. 1], Bressan [51], Godlewski and Raviart [138, pp. 1-104], Holden and Risebro [184, Chap. 5], LeFloch [213, Chap. VI]. The reader is referred to these references to acquire a deeper understanding of hyperbolic systems.

80.1 Weak solutions and examples

In this section, we introduce the concept of hyperbolic systems and weak solutions, and we give examples of hyperbolic systems.

80.1.1 First-order quasilinear hyperbolic systems

Let $m \in \mathbb{N}\backslash\{0\}$. Let $\mathcal{A}$ be a subset of $\mathbb{R}^m$ henceforth called *admissible set of states*. We use boldface notation for elements of $\mathcal{A}$ to emphasize the difference with scalar conservation equations. We keep the usual boldface notation for vectors in $\mathbb{R}^d$. Let $\mathbb{A}_l \in \mathrm{Lip}(\mathcal{A};\mathbb{R}^{m\times m})$ be some Lipschitz matrix-valued fields, for all $l \in \{1{:}d\}$. Let D be a polyhedron in $\mathbb{R}^d$, let $\boldsymbol{u}_0 \in \mathcal{A}$, and consider the following Cauchy problem:

$$\partial_t \boldsymbol{u} + \sum_{l\in\{1:d\}} \mathbb{A}_l(\boldsymbol{u})\partial_l \boldsymbol{u} = \boldsymbol{0}, \quad \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}), \quad (\boldsymbol{x},t) \in D\times\mathbb{R}_+, \tag{80.1}$$

with $\mathbb{R}_+ := [0,\infty)$. The dependent variable $\boldsymbol{u}$ is considered as a column vector $\boldsymbol{u} := (u_1,\dots,u_m)^\mathsf{T}$. Systems of PDEs like those in (80.1) are called *first-order quasilinear systems*. We avoid questions regarding boundary conditions by assuming that either periodic boundary conditions are enforced, or there is a compact subset $S \subsetneq D$ s.t. $\boldsymbol{u}_{0|D\backslash S}$ is constant over each connected component of $D\backslash S$ (there is only one connected component if $d \geq 2$), and there exists some $T > 0$ such that $\boldsymbol{u}(\boldsymbol{x},t) = \boldsymbol{u}_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial D$ and all $t \in [0,T]$. We refer the reader to Dubois and LeFloch [112] for more general boundary conditions.

Definition 80.1 (Hyperbolicity). (i) *We say that* (80.1) *is hyperbolic if the matrix* $\mathbb{A}(\boldsymbol{v},\boldsymbol{n}) := \sum_{l\in\{1:d\}} n_l\mathbb{A}_l(\boldsymbol{v})$ *is diagonalizable with real eigenvalues for all* $\boldsymbol{v} \in \mathcal{A}$ *and any unit vector* $\boldsymbol{n} :=$

$(n_1,\dots,n_d)^\mathsf{T} \in \mathbb{R}^d$. (ii) *The system is said to be strictly hyperbolic if all the eigenvalues are distinct.*

To motivate the above definition, suppose that $D := (0,2\pi)^d_{\mathrm{per}}$ is the periodic torus in $\mathbb{R}^d$. Assume also that $\mathbb{A}_l(\boldsymbol{v}) := \mathbb{A}_l$ for all $l \in \{1{:}d\}$ and all $\boldsymbol{v} \in \mathcal{A}$ and that there is some $\boldsymbol{k} := (k_1,\dots,k_d) \in \mathbb{Z}^d$ and $\mathbf{U}_{\boldsymbol{k}} \in \mathbb{R}^m$ such that $\boldsymbol{u}_0(\boldsymbol{x}) = \mathbf{U}_{\boldsymbol{k}} e^{\mathsf{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ with $\mathsf{i}^2 = -1$. Let us denote $\boldsymbol{n} := \frac{\boldsymbol{k}}{\|\boldsymbol{k}\|_{\ell^2}}$ and $\mathbb{A}(\boldsymbol{n}) := \frac{1}{\|\boldsymbol{k}\|_{\ell^2}} \sum_{l\in\{1:d\}} k_l \mathbb{A}_l$. If the system (80.1) is hyperbolic, Definition 80.1 implies that $\mathbb{A}(\boldsymbol{n})$ is diagonalizable in $\mathbb{R}$. Let $\lambda_1(\boldsymbol{k}),\dots,\lambda_m(\boldsymbol{k})$ be the eigenvalues of $\mathbb{A}(\boldsymbol{n})$ and $\mathbf{V}_1(\boldsymbol{k}),\dots,\mathbf{V}_m(\boldsymbol{k})$ be the associated unit eigenvectors. Then $\mathbf{U}_{\boldsymbol{k}}$ can be expanded in the eigenbasis as $\mathbf{U}_{\boldsymbol{k}} := \sum_{j\in\{1:m\}} \alpha_j(\boldsymbol{k})\mathbf{V}_j(\boldsymbol{k})$, so that $\boldsymbol{u}_0(\boldsymbol{x}) = \sum_{j\in\{1:m\}} \alpha_j(\boldsymbol{k}) e^{\mathsf{i}\boldsymbol{k}\cdot\boldsymbol{x}} \mathbf{V}_j(\boldsymbol{k})$. This leads to the following explicit representation of the solution to (80.1).

Lemma 80.2 (Plane wave solution). *Under the above assumptions, the unique solution to* (80.1) *is* $\boldsymbol{u}(\boldsymbol{x},t) = \sum_{j\in\{1:m\}} \alpha_j(\boldsymbol{k}) e^{\mathsf{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \lambda_j(\boldsymbol{k})\|\boldsymbol{k}\|_{\ell^2} t)} \mathbf{V}_j(\boldsymbol{k})$.

Proof. Let $(\boldsymbol{x},t) \in D\times\mathbb{R}_+$. We have

$$\begin{aligned}
\partial_t \boldsymbol{u}(\boldsymbol{x},t) &= -\mathsf{i} \sum_{j\in\{1:m\}} \lambda_j(\boldsymbol{k}) \|\boldsymbol{k}\|_{\ell^2} \alpha_j(\boldsymbol{k}) e^{\mathsf{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \lambda_j(\boldsymbol{k})\|\boldsymbol{k}\|_{\ell^2} t)} \mathbf{V}_j(\boldsymbol{k}),\\
\sum_{l\in\{1:m\}} \mathbb{A}_l \partial_l \boldsymbol{u}(\boldsymbol{x},t) &= \mathsf{i} \sum_{l\in\{1:m\}} k_l \mathbb{A}_l \sum_{j\in\{1:m\}} \alpha_j(\boldsymbol{k}) e^{\mathsf{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \lambda_j(\boldsymbol{k})\|\boldsymbol{k}\|_{\ell^2} t)} \mathbf{V}_j(\boldsymbol{k})\\
&= \mathsf{i} \sum_{j\in\{1:m\}} \lambda_j(\boldsymbol{k}) \|\boldsymbol{k}\|_{\ell^2} \alpha_j(\boldsymbol{k}) e^{\mathsf{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \lambda_j(\boldsymbol{k})\|\boldsymbol{k}\|_{\ell^2} t)} \mathbf{V}_j(\boldsymbol{k}),
\end{aligned}$$

so that $\partial_t \boldsymbol{u} + \sum_{l\in\{1:d\}} \mathbb{A}_l \partial_l \boldsymbol{u} = 0$ for all $(\boldsymbol{x},t) \in D\times\mathbb{R}_+$. Note also that $\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0$. Finally, the solution is unique since the system is linear. □

The above method generalizes to arbitrary initial data by using Fourier series techniques. The solution to the one-dimensional linear problem is investigated in Exercise 80.1.

Remark 80.3 (Change of variables). Let $\boldsymbol{\theta} : \mathcal{B} \subset \mathbb{R}^m \to \mathcal{A} \subset \mathbb{R}^m$ be a C^1-diffeomorphism and consider the change of variable $\boldsymbol{u} = \boldsymbol{\theta}(\boldsymbol{v})$. Then (80.1) can be rewritten

$$D\boldsymbol{\theta}(\boldsymbol{v})\partial_t \boldsymbol{v} + \sum_{l\in\{1:d\}} \mathbb{A}_l(\boldsymbol{\theta}(\boldsymbol{v}))(D\boldsymbol{\theta}(\boldsymbol{v})\partial_l \boldsymbol{v}) = \mathbf{0}.$$

After setting $\mathbb{B}_l(\boldsymbol{v}) := (D\boldsymbol{\theta}(\boldsymbol{v}))^{-1}\mathbb{A}_l(\boldsymbol{\theta}(\boldsymbol{v}))D\boldsymbol{\theta}(\boldsymbol{v})$, we conclude (at least informally) that (80.1) is equivalent to $\partial_t \boldsymbol{v} + \sum_{l\in\{1:d\}} \mathbb{B}_l(\boldsymbol{v})\partial_l \boldsymbol{v} = \mathbf{0}$. The matrices $\mathbb{B}(\boldsymbol{v},\boldsymbol{n}) := \sum_{l\in\{1:d\}} n_l \mathbb{B}_l(\boldsymbol{v})$ and $\mathbb{A}(\boldsymbol{v},\boldsymbol{n}) := \sum_{l\in\{1:d\}} n_l \mathbb{A}_l(\boldsymbol{v})$ being similar, this first-order quasilinear system is hyperbolic (or strictly hyperbolic) iff (80.1) is hyperbolic (or strictly hyperbolic). This shows that the notion of hyperbolicity is invariant under any smooth change of variables. □

80.1.2 Hyperbolic systems in conservative form

In the rest of this chapter, we are going to restrict our attention to first-order quasilinear systems that can be written in conservative form as follows:

$$\partial_t \boldsymbol{u} + \nabla\cdot\mathbb{f}(\boldsymbol{u}) = \mathbf{0}, \qquad \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}), \qquad (\boldsymbol{x},t) \in D\times\mathbb{R}_+. \tag{80.2}$$

The conserved variable $\boldsymbol{u}$ takes values in $\mathcal{A} \subset \mathbb{R}^m$ and the flux $\mathbb{f}$ is assumed to be s.t. $\mathbb{f} \in \mathrm{Lip}(\mathcal{A};\mathbb{R}^{m\times d})$. The set $\mathcal{A}$ is again called *admissible set of states*. For a generic element $\boldsymbol{v} \in \mathcal{A}$, the

flux is a matrix with entries $\mathbb{f}_{il}(\boldsymbol{v})$ for all $i \in \{1{:}m\}$ and all $l \in \{1{:}d\}$. In (80.2), $\nabla{\cdot}\mathbb{f}(\boldsymbol{u})$ evaluated at $(\boldsymbol{x},t)$ is a column vector in $\mathbb{R}^m$ with entries $(\nabla{\cdot}\mathbb{f}(\boldsymbol{u}))_i = \sum_{l\in\{1:d\}} \partial_{x_l}\mathbb{f}_{il}(\boldsymbol{u}(\boldsymbol{x},t))$, $\forall i \in \{1{:}m\}$.

For every unit vector $\boldsymbol{n} := (n_1 \dots, n_d)^\mathsf{T} \in \mathbb{R}^d$ and every state $\boldsymbol{v} \in \mathcal{A}$, we denote by $\mathbb{f}(\boldsymbol{v}){\cdot}\boldsymbol{n}$ the column vector in $\mathbb{R}^m$ with entries $\sum_{l\in\{1:d\}} n_l\mathbb{f}_{il}(\boldsymbol{v})$, $\forall i \in \{1{:}m\}$. Denoting by $\mathbb{A}_l(\boldsymbol{v}) \in \mathbb{R}^{m\times m}$ the matrix with entries $\partial_{v_j}\mathbb{f}_{il}(\boldsymbol{v})$, $\forall i,j \in \{1{:}m\}$, and assuming that the solution $\boldsymbol{u}$ is smooth, the conservation equation in (80.2) can be rewritten in the quasilinear form (80.1). Consistently with Definition 80.1, we say that (80.2) is hyperbolic iff the matrix $\mathbb{A}(\boldsymbol{v},\boldsymbol{n}) \in \mathbb{R}^{m\times m}$ with entries

$$\big(\mathbb{A}(\boldsymbol{v},\boldsymbol{n})\big)_{ij} := \sum_{l\in\{1:d\}} n_l\partial_{v_j}\mathbb{f}_{il}(\boldsymbol{v}), \qquad \forall i,j \in \{1{:}m\}, \tag{80.3}$$

is diagonalizable over $\mathbb{R}$ for all $\boldsymbol{v} \in \mathcal{A}$ and all unit vectors $\boldsymbol{n} \in \mathbb{R}^d$.

When (80.1) can be rewritten in the conservative form (80.2), we say that $\boldsymbol{u}$ is a conserved variable. There is a clear notion of weak solutions for the PDE (80.2) in conservative form. Recall that $C_0^1(D{\times}\mathbb{R}_+;\mathbb{R}^m)$ is composed of $\mathbb{R}^m$-valued functions that are compactly supported in $D{\times}\mathbb{R}_+$, and that these functions can be nonzero over a compact subset of $D{\times}\{t{=}0\}$.

Definition 80.4 (Weak solution). *We say that* $\boldsymbol{u} \in L^\infty_{\mathrm{loc}}(D{\times}\mathbb{R}_+;\mathbb{R}^m)$ *is a* weak solution *to* (80.2) *if for all* $\boldsymbol{\phi} \in C_0^1(D{\times}\mathbb{R}_+;\mathbb{R}^m)$*, we have*

$$\int_0^\infty \int_D \big(\boldsymbol{u}{\cdot}\partial_t\boldsymbol{\phi} + \mathbb{f}(\boldsymbol{u}){:}\nabla\boldsymbol{\phi}\big)\,\mathrm{d}x\mathrm{d}t + \int_D \boldsymbol{\phi}(\boldsymbol{x},0){\cdot}\boldsymbol{u}_0(\boldsymbol{x})\,\mathrm{d}x = 0, \tag{80.4}$$

where $\mathbb{f}(\boldsymbol{u}){:}\nabla\boldsymbol{\phi} := \sum_{i\in\{1:m\}}\sum_{l\in\{1:d\}} \mathbb{f}_{il}(\boldsymbol{u})\partial_l\phi_i$.

Giving a proper notion of weak solutions to (80.1) is far more delicate than for (80.2) since integration by parts in space is not possible. We refer the reader to Dal Maso et al. [98], Berthon et al. [31], where a suitable notion of weak solutions is proposed and the nonlinear stability of these solutions is investigated. Very much like for scalar conservation equations, there may be a nonuniqueness problem for the solution of (80.2) when $\mathbb{f}$ is nonlinear. One way to address this issue is to consider additional constraints like entropy inequalities.

Definition 80.5 (Entropy). *We say that* $(\eta,\boldsymbol{q})$ *is an entropy pair for* (80.2) *if the function* $\eta \in C^1(\mathcal{A};\mathbb{R})$ *is convex and if the function* $\boldsymbol{q} \in C^1(\mathcal{A};\mathbb{R}^d)$ *is such that*

$$\partial_{v_j} q_k(\boldsymbol{v}) = \sum_{i\in\{1:m\}} \partial_{v_i}\eta(\boldsymbol{v})\partial_{v_j}\mathbb{f}_{ik}(\boldsymbol{v}),$$

for all $j \in \{1{:}m\}$*, all* $k \in \{1{:}d\}$*, and all* $\boldsymbol{v} \in \mathcal{A}$*. The function* η *is called* entropy *and the function* $\boldsymbol{q}$ entropy flux.

Whenever an entropy pair $(\eta,\boldsymbol{q})$ is available for (80.2), one can select a physically relevant solution by enforcing entropy inequalities (nothing is said here about the uniqueness of such a solution). Specifically, one requests that the following holds true for all $\phi \in C_0^1(D{\times}\mathbb{R}_+;\mathbb{R}_+)$:

$$-\int_0^\infty \int_D (\eta(\boldsymbol{u})\partial_t\phi + \boldsymbol{q}(\boldsymbol{u}){\cdot}\nabla\phi)\,\mathrm{d}x\mathrm{d}t - \int_D \phi(\boldsymbol{x},0)\eta(\boldsymbol{u}_0)\,\mathrm{d}x \le 0. \tag{80.5}$$

Note that (80.5) implies that $\partial_t\eta(\boldsymbol{u}){+}\nabla{\cdot}\boldsymbol{q}(\boldsymbol{u}) \le 0$ in the sense of distributions in $D{\times}(0,\infty)$. Owing to the definition of the entropy flux, one readily verifies using the chain rule that if the weak solution is smooth, i.e., $\boldsymbol{u} \in C^1(D\times\mathbb{R}_+;\mathbb{R}^m)$, then the entropy inequalities (80.5) are actually equalities. An argument similar to that in the proof of Theorem 18.8 shows that the entropy inequalities are

equalities if $\boldsymbol{u}$ is piecewise smooth and continuous. The difficulty with the above entropy-based approach for systems is that, given an entropy, it is not clear whether an associated entropy flux exists, because the system of equations $\partial_{v_j} q_k(\boldsymbol{v}) = \sum_{i\in\{1:m\}} \partial_{v_i}\eta(\boldsymbol{v})\partial_{v_j}\mathbb{f}_{ik}(\boldsymbol{v})$ for all $j \in \{1{:}m\}$ and all $k \in \{1{:}d\}$ is in general overdetermined. We will see in the next section though that there are many physical examples of hyperbolic systems that admit nontrivial entropy pairs.

Another way to define a physically relevant solution to (80.2) is to consider the viscous regularization. For instance, given $\epsilon > 0$, the viscous regularization of (80.2) is the unique solution to the Cauchy problem

$$\partial_t \boldsymbol{u}_\epsilon + \nabla\cdot\mathbb{f}(\boldsymbol{u}_\epsilon) - \epsilon\Delta\boldsymbol{u}_\epsilon = \boldsymbol{0}, \qquad \boldsymbol{u}_\epsilon(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}), \quad (\boldsymbol{x},t)\in D\times\mathbb{R}_+. \tag{80.6}$$

We say that $\boldsymbol{u} \in L^\infty_{\mathrm{loc}}(D\times\mathbb{R}_+)$ is a vanishing viscosity solution to (80.2) over $D\times\mathbb{R}_+$ if $\|\boldsymbol{u}_\epsilon - \boldsymbol{u}\|_{L^1(D\times\mathbb{R}_+;\mathbb{R}^m)}$ as $\epsilon\to 0$. (Again, nothing is said here about the uniqueness of the solutions thus defined since these are delicate questions in general.) See Exercise 80.3 for a connection between the viscous regularization and entropy inequalities.

Remark 80.6 (Entropy inequality). Unlike for the scalar conservation equations, a general well-posedness theory for (80.2) is not available. Two major early works in this direction are the results by Lax [212, Thm. 9.1] and Glimm [136, Thm. 1.1] for one-dimensional hyperbolic systems, where it is shown that under some reasonable assumptions on the flux (see Theorem 80.18 for the details), and if the data satisfy some smallness conditions, then there exists a global weak solution to (80.2) that satisfies every entropy inequality for every entropy pair of the system. An important result by Bianchini and Bressan [32, Thm. 1] connects the vanishing-viscosity property with the entropy inequalities for this class of one-dimensional problems. The situation in higher dimensions is even worse. For instance, it is established in De Lellis and Székelyhidi [101], Chiodaroli et al. [81] that in two dimensions, one can construct initial data for the isentropic Euler equations for which there are infinitely many weak solutions that satisfy the entropy inequality associated with the physical entropy. The reader is also referred to Serre [251, Chap. 6] for a detailed analysis of the properties of (80.6). □

80.1.3 Examples

Example 80.7 (Scalar case). Assume that $m = 1$ and d is arbitrary, i.e., (80.2) is a scalar conservation equation. Let $\boldsymbol{n}\in\mathbb{R}^d$ be a unit vector so that $\mathbb{f}'(v)\cdot\boldsymbol{n}$ is a scalar for all v. This is a 1×1 matrix which is obviously diagonalizable with the unique real eigenvalue $\mathbb{f}'(v)\cdot\boldsymbol{n}$. Assuming that $\mathbb{f}\in\mathrm{Lip}(\mathbb{R};\mathbb{R})$, $\mathcal{A} := \mathbb{R}$ is an admissible set. □

Example 80.8 (Linear wave equation). Consider the linear system

$$\begin{cases} \partial_t u + \nabla\cdot\boldsymbol{v} = 0, \\ \partial_t \boldsymbol{v} + c^2\nabla u = 0, \end{cases} \qquad (\boldsymbol{x},t)\in\mathbb{R}^d\times\mathbb{R}_+, \tag{80.7}$$

where $c\neq 0$. Taking the time derivative of the first equation, the divergence of the second one and subtracting the results, we obtain the linear wave equation $\partial_{tt}u - c^2\Delta u = 0$. Using the notation $\boldsymbol{u} := (u,\boldsymbol{v}^\intercal)^\intercal$, we have $m = d+1$ and

$$\mathbb{f}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{v}^\intercal \\ c^2 u\mathbb{I}_d \end{pmatrix}, \quad \mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n} = \begin{pmatrix} \boldsymbol{v}\cdot\boldsymbol{n} \\ c^2 u\boldsymbol{n} \end{pmatrix}, \quad D(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = \begin{pmatrix} 0 & \boldsymbol{n}^\intercal \\ c^2\boldsymbol{n} & \mathbb{O}_d \end{pmatrix}.$$

One can verify (see Exercise 80.2) that the $(d+1)$ eigenpairs of the matrix $D(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n})$ are $(c,(1,c\boldsymbol{n}^\intercal)^\intercal)$, $(-c,(1,-c\boldsymbol{n}^\intercal)^\intercal)$, and $(0,(0,\boldsymbol{v}_l^\intercal)^\intercal)$ for all $l\in\{1{:}d{-}1\}$, where the vectors $\{\boldsymbol{v}_l\}_{l\in\{1:d-1\}}$ are such that $\{\boldsymbol{n},\boldsymbol{v}_1,\dots,\boldsymbol{v}_{d-1}\}$ forms an orthonormal basis of $\mathbb{R}^d$. An admissible set is $\mathcal{A} := \mathbb{R}\times\mathbb{R}^d$. □

Example 80.9 (p-system). The one-dimensional motion of an isentropic gas in Lagrangian coordinates is modeled by the p-system:

$$\begin{cases} \partial_t v - \partial_x u = 0, \\ \partial_t u + \partial_x p(v) = 0, \end{cases} \qquad (x,t) \in \mathbb{R}\times\mathbb{R}_+. \tag{80.8}$$

Here, we have $d = 1$ and $m = 2$. The dependent variables are the velocity u and the specific volume v. The map $v \mapsto p(v)$ is the pressure and is assumed to be in $C^2(\mathbb{R}_+;\mathbb{R})$ and to satisfy $0 < p''$ and $p' < 0$. A typical example is the γ-law, $p(v) := rv^{-\gamma}$ with $r > 0$ and $\gamma \geq 1$. Using the notation $\boldsymbol{u} := (v,u)^\mathsf{T}$, $\boldsymbol{n} := \pm\boldsymbol{e}_x$, we have

$$\mathbb{f}(\boldsymbol{u}) := \begin{pmatrix} -u \\ p(v) \end{pmatrix} \boldsymbol{e}_x, \qquad D(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = \boldsymbol{e}_x\cdot\boldsymbol{n} \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix}.$$

The system is hyperbolic owing to the assumption $p'(v) < 0$. The two eigenpairs of the matrix $D(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n})$ are $(\mp\sqrt{-p'(v)}, (1, \pm\sqrt{-p'(v)})^\mathsf{T})$. The reader is invited to verify that $\eta(\boldsymbol{u}) := \frac{1}{2}u^2 - P(v)$, where $P(v)$ is a primitive of $p(v)$, i.e., $P'(v) = p(v)$, is an entropy and the associated entropy flux is $\boldsymbol{q}(\boldsymbol{u}) := p(v)u\boldsymbol{e}_x$. The system is strictly hyperbolic. An admissible set is $\mathcal{A} := (0,\infty)\times\mathbb{R}$. □

Example 80.10 (Euler equations). Consider the Euler equations in $\mathbb{R}^d$:

$$\partial_t \boldsymbol{u} + \nabla\cdot(\mathbb{f}(\boldsymbol{u})) = \boldsymbol{0}, \quad \boldsymbol{u} := \begin{pmatrix} \rho \\ \boldsymbol{m} \\ E \end{pmatrix}, \quad \mathbb{f}(\boldsymbol{u}) := \begin{pmatrix} \boldsymbol{m}^\mathsf{T} \\ \boldsymbol{m}\otimes\frac{\boldsymbol{m}}{\rho} + p\mathbb{I}_d \\ \frac{\boldsymbol{m}^\mathsf{T}}{\rho}(E+p) \end{pmatrix}, \tag{80.9}$$

where ρ is the density, $\boldsymbol{m}$ the momentum (column vector), and E the total energy. An admissible set of states is $\mathcal{A} := \{(\rho,\boldsymbol{m},E) \mid \rho > 0,\ E - \frac{1}{2}\boldsymbol{m}^2/\rho > 0\}$. The pressure, p, is given by the equation of state which we assume to derive from a specific entropy, $\sigma(\tau,e)$, defined by the thermodynamics identity $T\,\mathrm{d}\sigma := \mathrm{d}e + p\,\mathrm{d}\tau$, where $\tau := \rho^{-1}$, $e := \rho^{-1}E - \frac{1}{2}\boldsymbol{v}^2$ is the *specific internal energy*, $\boldsymbol{v} := \rho^{-1}\boldsymbol{m}$ is the velocity of the fluid particles, and T is the temperature. Note that $\tau > 0$ and $e > 0$ for every admissible state, i.e., $\sigma : (0,\infty)^2 \to \mathbb{R}$. There are two key structural properties coming from thermodynamics, namely that $T > 0$ and that the function σ is strictly concave on $(0,\infty)^2$. The above thermodynamics identity means that $\partial_e\sigma = T^{-1}$ and $\partial_\tau\sigma = pT^{-1}$. These relations allow one to define T and p as functions of (τ,e). In the continuum mechanics literature, one uses ρ rather than τ, i.e., one considers the function $s : (0,\infty)^2 \to \mathbb{R}$ s.t. $s(\rho,e) := \sigma(\tau,e)$, and the above thermodynamics identity is written as $T\,\mathrm{d}s = \mathrm{d}e - p\rho^{-2}\,\mathrm{d}\rho$, up to an abuse of notation since T and p are now viewed as functions of (ρ,e). The equation of state defining the pressure then takes the form

$$p(\rho,e) = -\rho^2\frac{\partial_\rho s(\rho,e)}{\partial_e s(\rho,e)}. \tag{80.10}$$

For instance, one has $s(\rho,e) := \ln(e^{\frac{1}{\gamma-1}}\rho^{-1})$ for a *polytropic ideal gas*, so that $p(\rho,e) = (\gamma-1)\rho e$. The function s is called *specific entropy* or physical specific entropy and $-s$ mathematical specific entropy. Note though that s is not an entropy in the sense of Definition 80.5 since s is a function of two variables (density and specific internal energy) which are not the conserved variables. The key observation is that, after a change of variables, letting $\Phi : \mathcal{A} \to \mathbb{R}$ be s.t. $\Phi(\boldsymbol{u}) := s(\rho,\rho^{-1}E - \frac{1}{2}\rho^{-2}\boldsymbol{m}^2)$, the function $S : \mathcal{A} \to \mathbb{R}$ s.t. $S(\boldsymbol{u}) := -\rho\Phi(\boldsymbol{u})$ is an entropy in the sense of Definition 80.5; see Exercise 80.5.

Owing to thermodynamics, the change of variable $(\tau,e) \to (p,T^{-1})$ is bijective (the determinant of the Jacobian matrix is indeed equal to $\big(\partial_{\tau\tau}\sigma\partial_{ee}\sigma - (\partial_{\tau e}\sigma)^2\big)/\partial_e\sigma$, and this quantity

is negative since σ is strictly convex and $(\partial_e\sigma)^{-1} := T > 0$). This shows that we can set $\mathfrak{s}(p,T) := \sigma(\tau(p,T),e(p,T))$ and, up to an abuse of notation, we can define the specific heat at constant pressure $c_p(\rho,e) := T\partial_T\mathfrak{s}(p,T)$. Let $f \in C^2(\mathbb{R};\mathbb{R})$ be such that

$$f'(s) > 0, \qquad f'(s)c_p^{-1} - f''(s) > 0, \qquad \forall (\rho,e) \in (0,\infty)^2, \tag{80.11}$$

where $f'(s)$ and $f''(s)$ stand for $f'(s(\rho,e))$ and $f''(s(\rho,e))$. It is shown in Harten et al. [180] that the function $-\rho f(\Phi(\boldsymbol{u})) : \mathcal{A} \to \mathbb{R}$ is an entropy for (80.9) and the associated entropy flux is $-\boldsymbol{m}f(\Phi(\boldsymbol{u}))$. The reader is also referred to Guermond and Popov [155] for other details.

Let us abuse the notation by saying that p is now a function of ρ and s (this is legitimate owing to the implicit function theorem since $\partial_e s \neq 0$). Then it is shown in Exercise 80.5 that the concavity of σ and the condition $(\partial_e s)^{-1} > 0$ imply that $\partial_\rho p(\rho,s) > 0$. The quantity $c(\rho,s) := \sqrt{\partial_\rho p(\rho,s)}$ is called *sound speed*. We refer the reader to Godunov [139], Friedrichs and Lax [132], Harten et al. [180], Godlewski and Raviart [138, pp. 99-104] for further details on this question.

We now use Remark 80.3 to establish that (80.9) is hyperbolic. We make the change of variables $(\rho,\boldsymbol{m},E) \to (\rho,\boldsymbol{v},s)$ and assume that all the quantities that we manipulate are smooth with respect to space and time. Using the mass conservation equation, the momentum equation can be rewritten $\partial_t\boldsymbol{v} + \boldsymbol{v}\cdot\nabla\boldsymbol{v} + \frac{1}{\rho}\nabla p = 0$. Multiplying this equation by $\boldsymbol{v}$ gives $\partial_t(\frac{1}{2}\boldsymbol{v}^2) + \boldsymbol{v}\cdot\nabla(\frac{1}{2}\boldsymbol{v}^2) + \frac{1}{\rho}\boldsymbol{v}\cdot\nabla p = 0$. Subtracting this equation from $\partial_t\mathcal{E} + \boldsymbol{v}\cdot\nabla\mathcal{E} + \frac{1}{\rho}\nabla\cdot(p\boldsymbol{v}) = 0$ where $\mathcal{E} := E/\rho$, we obtain $\partial_t e + \boldsymbol{v}\cdot\nabla e + \frac{p}{\rho}\nabla\cdot\boldsymbol{v} = 0$. Moreover, multiplying the mass conservation equation by $\partial_\rho s$, multiplying the balance of specific internal energy by $\partial_e s$, adding the two results, and using the equation of state $p\partial_e s + \rho^2\partial_\rho s = 0$, we obtain the balance of specific entropy $\partial_t s + \boldsymbol{v}\cdot\nabla s = 0$. In conclusion, we have shown that (80.9) can be put into the form of the following first-order quasilinear system:

$$\partial_t\rho + \boldsymbol{v}\cdot\nabla\rho + \rho\nabla\cdot\boldsymbol{v} = 0, \qquad \partial_t\boldsymbol{v} + \boldsymbol{v}\cdot\nabla\boldsymbol{v} + \frac{1}{\rho}\nabla p = \boldsymbol{0}, \qquad \partial_t s + \boldsymbol{v}\cdot\nabla s = 0,$$

which can be recast into the form (80.1) by setting

$$\mathbb{A}_l(\boldsymbol{u}) := \begin{pmatrix} v_l & \rho\boldsymbol{e}_l^\intercal & 0 \\ \frac{\partial_\rho p}{\rho}\boldsymbol{e}_l & v_l\mathbb{I}_d & \frac{\partial_s p}{\rho}\boldsymbol{e}_l \\ 0 & \boldsymbol{0}^\intercal & v_l \end{pmatrix}, \quad \mathbb{A}(\boldsymbol{u},\boldsymbol{n}) = \begin{pmatrix} \boldsymbol{v}\cdot\boldsymbol{n} & \rho\boldsymbol{n}^\intercal & 0 \\ \frac{\partial_\rho p}{\rho}\boldsymbol{n} & \boldsymbol{v}\cdot\boldsymbol{n}\mathbb{I}_d & \frac{\partial_s p}{\rho}\boldsymbol{n} \\ 0 & \boldsymbol{0}^\intercal & \boldsymbol{v}\cdot\boldsymbol{n} \end{pmatrix}, \tag{80.12}$$

where $(\boldsymbol{e}_l)_{l\in\{1:d\}}$ is the canonical basis of $\mathbb{R}^d$. The reader is invited to verify that $(\boldsymbol{v}\cdot\boldsymbol{n} \mp \sqrt{\partial_\rho p(\rho,s)}, (\rho, \mp\sqrt{\partial_\rho p(\rho,s)}\boldsymbol{n}^\intercal, 0)^\intercal)$ are eigenpairs of multiplicity 1. Let $\{\mathbf{V}_1,\dots,\mathbf{V}_{d-1}\}$ be such that $\{\boldsymbol{n},\mathbf{V}_1,\dots,\mathbf{V}_{d-1}\}$ forms an orthonormal basis of $\mathbb{R}^d$. Then $(\boldsymbol{v}\cdot\boldsymbol{n}, (-\partial_s p(\rho,s), \mathbf{V}_l^\intercal, \partial_\rho p(\rho,s))^\intercal)$ is an eigenpair for all $l \in \{1{:}d{-}1\}$, that is, the multiplicity of the eigenvalue $\boldsymbol{v}\cdot\boldsymbol{n}$ is $(d-1)$. □

80.2 Riemann problem

A theory for the well-posedness of (80.2) is not available for general fluxes and data, but there is a clear notion of solution to the Riemann problem. The purpose of this section is to present some elementary facts about this problem. Given a pair of states $(\boldsymbol{u}_L,\boldsymbol{u}_R) \in \mathcal{A}\times\mathcal{A}$ and a unit vector $\boldsymbol{n} \in \mathbb{R}^d$, we consider the following one-dimensional *Riemann problem*:

$$\partial_t\boldsymbol{u} + \partial_x(\mathbb{f}(\boldsymbol{u})\cdot\boldsymbol{n}) = 0, \qquad \boldsymbol{u}(x,0) := \begin{cases} \boldsymbol{u}_L & \text{if } x \le 0, \\ \boldsymbol{u}_R & \text{if } 0 < x. \end{cases} \tag{80.13}$$

Further assumptions on the data will be made when appropriate.

80.2.1 Expansion wave, contact discontinuity, and shock

The goal of this section is to build some elementary weak solutions to the Riemann problem (80.13). These weak solutions will be patched together in the next section to construct a *vanishing-viscosity* solution to (80.13).

Recall the matrix $\mathbb{A}(\boldsymbol{v},\boldsymbol{n})$ with entries defined in (80.3) for all $\boldsymbol{v}\in\mathcal{A}$ and every unit vector $\boldsymbol{n}\in\mathbb{R}^d$. Let $(\lambda_l(\boldsymbol{v}),\boldsymbol{r}_l(\boldsymbol{v}))\in\mathbb{R}\times\mathbb{R}^m$ for all $l\in\{1{:}m\}$, be the eigenpairs of $\mathbb{A}(\boldsymbol{v},\boldsymbol{n})$ with the convention that $\lambda_1(\boldsymbol{v})\le\ldots\le\lambda_m(\boldsymbol{v})$. (We omit the dependence on $\boldsymbol{n}$ to simplify the notation.) We assume that the dependence of the eigenpairs with respect to $\boldsymbol{v}$ is at least of class C^1. The eigenpairs are often called *characteristic families* in the literature. The eigenvectors are normalized in some way which will be specified later.

Definition 80.11 (Genuinely nonlinear, linearly degenerate eigenpairs). *Let $l\in\{1{:}m\}$ and $\boldsymbol{n}\in\mathbb{R}^d$ be a unit vector. We say that the l-th eigenpair is* genuinely nonlinear *if $D\lambda_l(\boldsymbol{v})\cdot\boldsymbol{r}_l(\boldsymbol{v})\neq 0$ for all $\boldsymbol{v}\in\mathcal{A}$, and that it is* linearly degenerate *if $D\lambda_l(\boldsymbol{v})\cdot\boldsymbol{r}_l(\boldsymbol{v})=0$ for all $\boldsymbol{v}\in\mathcal{A}$. Here, $D\lambda_l(\boldsymbol{v})$ is viewed as a column vector in $\mathbb{R}^m$.*

Example 80.12 (Scalar conservation). When $m=1$, $\lambda(v)=f'(v)$ is the only eigenvalue. The eigenpair $(\lambda(v),\boldsymbol{e}_x)$ is genuinely nonlinear iff $f''(v)\neq 0$ for all $v\in\mathbb{R}$, which is the case if f is either strictly convex or strictly concave, and it is linearly degenerate iff $f''(v)=0$ for all $v\in\mathbb{R}$, i.e., iff $f(v)=av$, $a\in\mathbb{R}$. Note that it is possible that the eigenpair is neither genuinely nonlinear nor linearly degenerate. This situation is more difficult to handle. □

Example 80.13 (Euler equations). Recalling Example 80.10 and using the dependent variable $(\rho,\boldsymbol{v}^\mathsf{T},s)^\mathsf{T}$, we obtain $D\lambda_1\cdot\boldsymbol{r}_1=-c^{-1}(\frac{\rho}{2}\partial_{\rho\rho}p+\partial_\rho p)$, $D\lambda_l\cdot\boldsymbol{r}_l=0$, $\forall l\in\{2{:}d\}$, $D\lambda_{d+1}\cdot\boldsymbol{r}_{d+1}=c^{-1}(\frac{\rho}{2}\partial_{\rho\rho}p+\partial_\rho p)$, where $c(\rho,s):=\sqrt{\partial_\rho p(\rho,s)}$ is the sound speed. Up to an abuse of notation, we have $\frac{\rho}{2}\partial_{\rho\rho}p(\rho,s)+\partial_\rho p(\rho,s)=\frac{1}{2\rho^3}\partial_{\tau\tau}p(\tau,s)$. Hence, assuming $\partial_{\tau\tau}p(\tau,s)>0$, which is the case for many realistic fluids, the first and the last eigenvalues are genuinely nonlinear. For instance, for the γ-law where $s(\rho,e):=\ln(e^{\frac{1}{\gamma-1}}\rho^{-1})$, we have $p(\rho,e)=(\gamma-1)\rho e$ and $p(\tau,s)=(\gamma-1)\tau^{-\gamma}e^{(\gamma-1)s}$, so that $\partial_{\tau\tau}p(\tau,s)=(\gamma+1)\gamma(\gamma-1)\tau^{-\gamma-2}e^{(\gamma-1)s}>0$ for all $(\tau,s)\in(0,\infty)\times\mathbb{R}$. Finally, the eigenpairs for all $l\in\{2{:}d\}$ are linearly degenerate. □

Let us assume for the sake of simplicity that the eigenpairs are either genuinely nonlinear or linearly degenerate. Let us normalize the eigenvectors in such a way that $\|\boldsymbol{r}_l(\boldsymbol{v})\|_{\ell^2}=1$ if the l-th eigenpair is linearly degenerate and $D\lambda_l(\boldsymbol{v})\cdot\boldsymbol{r}_l(\boldsymbol{v})=1$ if the l-th eigenpair is genuinely nonlinear. Let us first look for a self-similar solution to (80.13) in the form $\boldsymbol{u}(x,t)=\boldsymbol{w}(\frac{x}{t})$ for some smooth function $\boldsymbol{w}$. Setting $\xi:=\frac{x}{t}$ and using the chain rule, we see that $\boldsymbol{u}(x,t)=\boldsymbol{w}(\xi)$ solves (80.13) if and only if

$$\mathbb{A}(\boldsymbol{w}(\xi),\boldsymbol{n})\boldsymbol{w}'(\xi)=\xi\boldsymbol{w}'(\xi). \tag{80.14}$$

This is possible only if either $\boldsymbol{w}$ is constant or $(\xi,\boldsymbol{w}'(\xi))$ is an eigenpair of $\mathbb{A}(\boldsymbol{w}(\xi),\boldsymbol{n})$. If $(\xi,\boldsymbol{w}'(\xi))$ is an eigenpair, there is $l\in\{1{:}m\}$ such that $\lambda_l(\boldsymbol{w}(\xi))=\xi$ and $\boldsymbol{w}'(\xi)$ is proportional to $\boldsymbol{r}_l(\boldsymbol{w}(\xi))$, i.e., there is $\gamma(\xi)\in\mathbb{R}$ s.t. $\boldsymbol{w}'(\xi)=\gamma(\xi)\boldsymbol{r}_l(\boldsymbol{w}(\xi))$. Let us assume that the l-th eigenpair is genuinely nonlinear. Then differentiating $\lambda_l(\boldsymbol{w}(\xi))=\xi$ with respect to ξ, we obtain $\gamma(\xi)D\lambda_l(\boldsymbol{w}(\xi))\cdot\boldsymbol{r}_l(\boldsymbol{w}(\xi))=1$, i.e., $\gamma(\xi)=1$ owing to the adopted normalization. Hence, $\boldsymbol{w}$ satisfies $\boldsymbol{w}'(\xi)=\boldsymbol{r}_l(\boldsymbol{w}(\xi))$ if the l-th eigenpair is genuinely nonlinear. Let us now construct a particular weak solution to (80.13) which we call *expansion wave*.

Lemma 80.14 (Expansion wave). *Suppose that the l-th eigenpair is genuinely nonlinear. Let $\boldsymbol{u}_Z\in\mathcal{A}$ and let $\xi_Z:=\lambda_l(\boldsymbol{u}_Z)$. Let $\delta^*>0$ be such that $\boldsymbol{w}\in C^1((\xi_Z-\delta^*,\xi_Z+\delta^*);\mathbb{R}^m)$ solves the ordinary differential equation $\boldsymbol{w}'(\xi)=\boldsymbol{r}_l(\boldsymbol{w}(\xi))$ with $\boldsymbol{w}(\xi_Z)=\boldsymbol{u}_Z$ (this is legitimate since we assumed that $\boldsymbol{r}_l(\boldsymbol{v})$ and $\lambda_l(\boldsymbol{v})$ depend smoothly on $\boldsymbol{v}$).* (i) *The identity $\lambda_l(\boldsymbol{w}(\xi))=\xi$ holds for all*

$\xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*)$. (ii) *Let* $\xi_L \in (\xi_Z - \delta^*, \xi_Z)$ *and set* $\boldsymbol{u}_L := \boldsymbol{w}(\xi_L)$. *Let* $\xi_R \in (\xi_Z, \xi_Z + \delta^*)$ *and set* $\boldsymbol{u}_R := \boldsymbol{w}(\xi_R)$. *Then* $\lambda_l(u_L) < \lambda_l(u_R)$, *and the function*

$$\boldsymbol{u}(x,t) := \begin{cases} \boldsymbol{u}_L & \text{if } \frac{x}{t} \le \lambda_l(\boldsymbol{u}_L), \\ \boldsymbol{w}(\xi) & \text{if } \lambda_l(\boldsymbol{u}_L) < \frac{x}{t} \le \lambda_l(\boldsymbol{u}_R), \\ \boldsymbol{u}_R & \text{if } \lambda_l(\boldsymbol{u}_R) < \frac{x}{t}, \end{cases} \tag{80.15}$$

is a self-similar weak solution to (80.13).

Proof. (i) Since $\boldsymbol{w}'(\xi) = \boldsymbol{r}_l(\boldsymbol{w}(\xi))$, we have $\mathbb{A}(\boldsymbol{w}(\xi), \boldsymbol{n})\boldsymbol{w}'(\xi) = \lambda_l(\boldsymbol{w}(\xi))\boldsymbol{w}'(\xi)$ for all $\xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*)$. Owing to the normalization of $\boldsymbol{r}_l$, we infer that

$$\frac{\mathrm{d}}{\mathrm{d}\xi}(\xi - \lambda_l(\boldsymbol{w}(\xi))) = 1 - D\lambda_l(\boldsymbol{w}(\xi))\cdot\boldsymbol{r}_l(\boldsymbol{w}(\xi)) = 0.$$

Hence, $\xi - \lambda_l(\boldsymbol{w}(\xi))$ is a constant function in ξ, and evaluating this function at $\xi = \xi_Z$, we obtain $\xi - \lambda_l(\boldsymbol{w}(\xi)) = \xi_Z - \lambda_l(\boldsymbol{w}(\xi_Z)) = \xi_Z - \lambda_l(\boldsymbol{u}_Z) = 0$, so that

$$\xi = \lambda_l(\boldsymbol{w}(\xi)), \qquad \forall \xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*).$$

(ii) Since $\xi_L \in (\xi_Z - \delta^*, \xi_Z)$, we infer that $\xi_L = \lambda_l(\boldsymbol{w}(\xi_L))$, and by definition of $\boldsymbol{u}_L$, this means that $\xi_L = \lambda_l(\boldsymbol{u}_L)$. Similarly, we have $\xi_R = \lambda_l(\boldsymbol{u}_R)$. This implies that $\lambda_l(u_L) = \xi_L < \xi_Z < \xi_R = \lambda_l(u_R)$. Moreover, the above identities prove that the function $\boldsymbol{u}$ defined in (80.15) is continuous. Since $\boldsymbol{u}$ is piecewise smooth, an argument similar to that invoked in the proof of Theorem 18.8 shows that $\boldsymbol{u}$ is a weak solution iff it satisfies $\partial_t \boldsymbol{u} + \mathbb{A}(\boldsymbol{u}, \boldsymbol{n})\partial_x \boldsymbol{u} = \boldsymbol{0}$ in the three (open) angular sectors $\{\frac{x}{t} < \lambda_l(\boldsymbol{u}_L)\}$, $\{\lambda_l(\boldsymbol{u}_L) < \frac{x}{t} < \lambda_l(\boldsymbol{u}_R)\}$, and $\{\lambda_l(\boldsymbol{u}_R) < \frac{x}{t}\}$. The claim trivially holds true in the first and third sectors where $\boldsymbol{u}$ is constant, and in the second sector, the claim follows from $\mathbb{A}(\boldsymbol{w}(\xi), \boldsymbol{n})\boldsymbol{w}'(\xi) = \lambda_l(\boldsymbol{w}(\xi))\boldsymbol{w}'(\xi) = \xi\boldsymbol{w}'(\xi)$ since (80.14) is satisfied in the second sector. □

There are also solutions associated with the linearly degenerate eigenpairs. These solutions, called *contact discontinuities*, are piecewise constants separated by a discontinuity moving at some speed s.

Lemma 80.15 (Rankine–Hugoniot). *Let* $s \in \mathbb{R}$. *The function*

$$\boldsymbol{u}(x,t) := \begin{cases} \boldsymbol{u}_L & \text{if } \frac{x}{t} \le s, \\ \boldsymbol{u}_R & \text{if } s < \frac{x}{t}, \end{cases} \tag{80.16}$$

is a weak solution to (80.13) *if and only if the speed* s *is s.t. the following* Rankine–Hugoniot *condition holds true:*

$$\mathbb{f}(\boldsymbol{u}_L)\cdot\boldsymbol{n} - \mathbb{f}(\boldsymbol{u}_R)\cdot\boldsymbol{n} = s(\boldsymbol{u}_L - \boldsymbol{u}_R). \tag{80.17}$$

Proof. Integrate (80.13) over $(-1,1)\times(0,t)$ with $t \le \frac{1}{s}$ and use (80.16). □

The Rankine–Hugoniot condition is a necessary and sufficient compatibility condition expressing that (80.16) is indeed a weak solution to (80.13).

Lemma 80.16 (Contact discontinuity). *Assume that the* l*-th eigenpair is linearly degenerate. Let* $\boldsymbol{u}_Z \in \mathcal{A}$ *and set* $\xi_Z := \lambda_l(\boldsymbol{u}_Z)$. *Let* $\delta^* > 0$ *be such that* $\boldsymbol{z} \in C^1((\xi_Z - \delta^*, \xi_Z + \delta^*); \mathbb{R}^m)$ *solves the ordinary differential equation* $\boldsymbol{z}'(\xi) = \boldsymbol{r}_l(\boldsymbol{z}(\xi))$ *with* $\boldsymbol{z}(\xi_Z) = \boldsymbol{u}_Z$. (i) *The identity* $\lambda_l(\boldsymbol{z}(\xi)) = \lambda_l(\boldsymbol{u}_Z)$

holds for all $\xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*)$. (ii) *Let* $\xi_L \in (\xi_Z - \delta^*, \xi_Z)$ *and set* $\boldsymbol{u}_L := \boldsymbol{z}(\xi_L)$. *Let* $\xi_R \in (\xi_Z, \xi_Z + \delta^*)$ *and set* $\boldsymbol{u}_R := \boldsymbol{z}(\xi_R)$. *Then the function defined by*

$$\boldsymbol{u}(x,t) := \begin{cases} \boldsymbol{u}_L & \text{if } \frac{x}{t} \le \lambda_l(\boldsymbol{u}_Z), \\ \boldsymbol{u}_R & \text{if } \lambda_l(\boldsymbol{u}_Z) < \frac{x}{t}, \end{cases} \tag{80.18}$$

is a self-similar weak solution to (80.13).

Proof. (i) We have $\frac{\mathrm{d}}{\mathrm{d}\xi}\lambda_l(\boldsymbol{z}(\xi)) = D\lambda_l(\boldsymbol{z}(\xi))\cdot\boldsymbol{z}'(\xi) = D\lambda_l(\boldsymbol{z}(\xi))\cdot\boldsymbol{r}_l(\boldsymbol{z}(\xi)) = 0$. Hence, $\lambda_l(\boldsymbol{z}(\xi)) = \lambda_l(\boldsymbol{z}(\xi_Z)) = \lambda_l(\boldsymbol{u}_Z)$ for all $\xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*)$.
(ii) Recalling that we use the notation $D\mathbb{f}(\boldsymbol{z}(\xi))\cdot\boldsymbol{n} = \mathbb{A}(\boldsymbol{z}(\xi), \boldsymbol{n})$, the following argument shows that the Rankine–Hugoniot condition holds true:

$$\begin{aligned} \big(\mathbb{f}(\boldsymbol{u}_R) - \mathbb{f}(\boldsymbol{u}_L)\big)\cdot\boldsymbol{n} &= \int_{\xi_L}^{\xi_R} \frac{\mathrm{d}}{\mathrm{d}\xi}\mathbb{f}(\boldsymbol{z}(\xi))\cdot\boldsymbol{n}\,\mathrm{d}\xi = \int_{\xi_L}^{\xi_R} D\mathbb{f}(\boldsymbol{z}(\xi))\cdot\boldsymbol{n}\,\boldsymbol{z}'(\xi)\,\mathrm{d}\xi \\ &= \int_{\xi_L}^{\xi_R} \mathbb{A}(\boldsymbol{z}(\xi),\boldsymbol{n})\,\boldsymbol{r}_l(\boldsymbol{z}(\xi))\,\mathrm{d}\xi = \int_{\xi_L}^{\xi_R} \lambda_l(\boldsymbol{z}(\xi))\boldsymbol{r}_l(\boldsymbol{z}(\xi))\,\mathrm{d}\xi \\ &= \lambda_l(\boldsymbol{u}_Z)\int_{\xi_L}^{\xi_R} \boldsymbol{z}'(\xi)\,\mathrm{d}\xi = \lambda_l(\boldsymbol{u}_Z)(\boldsymbol{u}_R - \boldsymbol{u}_L). \end{aligned}$$

We conclude the proof by invoking Lemma 80.15. □

We finish the discussion with a third class of waves that are called *shocks*. We are not going to go through the construction of these waves since they involve lengthy arguments invoking the implicit function theorem which are tangential to the objectives of the book. The essential result is the following.

Lemma 80.17 (Shock). *Let* $\boldsymbol{u}_Z \in \mathcal{A}$, *assume that the eigenvalue* $\lambda_l(\boldsymbol{u}_Z)$ *has multiplicity 1, and let* $\xi_Z := \lambda_l(\boldsymbol{u}_Z)$. (i) *There exists* $\delta^* > 0$ *and functions* $s_l \in C^0((\xi_Z - \delta^*, \xi_Z + \delta^*); \mathbb{R})$, $\boldsymbol{z}_l \in C^0((\xi_Z - \delta^*, \xi_Z + \delta^*); \mathbb{R}^m)$ *s.t.*

$$\big(\mathbb{f}(\boldsymbol{z}_l(\xi)) - \mathbb{f}(\boldsymbol{u}_Z)\big)\cdot\boldsymbol{n} = s_l(\xi)(\boldsymbol{z}_l(\xi) - \boldsymbol{u}_Z), \quad \forall \xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*). \tag{80.19}$$

(ii) *Let us fix* $\xi \in (\xi_Z - \delta^*, \xi_Z + \delta^*)$ *and set* $s := s_l(\xi)$. *If* $\xi < \xi_Z$, *set* $\boldsymbol{u}_L := \boldsymbol{z}_l(\xi)$ *and* $\boldsymbol{u}_R := \boldsymbol{u}_Z$, *whereas if* $\xi_Z < \xi$, *set* $\boldsymbol{u}_L := \boldsymbol{u}_Z$ *and* $\boldsymbol{u}_R := \boldsymbol{z}_l(\xi)$. *Then the function defined by* $\boldsymbol{u}(x,t) := \boldsymbol{u}_L$ *if* $\frac{x}{t} \le s$ *and* $\boldsymbol{u}(x,t) := \boldsymbol{u}_R$ *if* $s < \frac{x}{t}$ *is a self-similar weak solution to* (80.13). *(This solution is called* l*-shock if the* l*-th eigenpair is genuinely nonlinear.)*

Proof. See Holden and Risebro [184, Thm. 5.11] or Godlewski and Raviart [138, Thm. I.4.1]. □

Let us now return to the Riemann problem (80.13). Given a pair of states $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A}^2$, the hard question consists of piecing together all the above elementary solutions so as to form one weak solution to the Riemann problem (80.13) that is physically relevant, i.e., that is a vanishing-viscosity solution. An answer to this question is available if the states $(\boldsymbol{u}_L, \boldsymbol{u}_R)$ are close enough.

Theorem 80.18 (Lax). *Assume that* (80.13) *is strictly hyperbolic (i.e., all the eigenvalues are real with multiplicity 1) and that for all* $l \in \{1{:}m\}$, *the* l*-th eigenpair is either genuinely nonlinear or linearly degenerate. Then there exists* $\delta > 0$ *such that for every pair* $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A}^2$ *satisfying* $\|\boldsymbol{u}_L - \boldsymbol{u}_R\|_{\ell^2(\mathbb{R}^m)} \le \delta$, *the Riemann problem* (80.13) *has a weak solution that consists of at most* $(m+1)$ *constant states separated by expansion waves, shocks or contact discontinuities, and this solution is a vanishing-viscosity solution.*

Proof. See Lax [212, Thm. 9.1]. The vanishing-viscosity property is established in Bianchini and Bressan [32, Thm. 1]. □

Theorem 80.18 says that there are $2m$ numbers $\{\lambda_l^\pm(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)\}_{l\in\{1:m\}}$ such that (the dependency on $(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ is omitted for simplicity)

$$\lambda_1^- \le \lambda_1^+ \le \lambda_2^- \le \ldots \le \lambda_m^- \le \lambda_m^+, \tag{80.20}$$

and these numbers define up to $(2m+1)$ sectors in the (x,t) plane (some could be reduced to a line), $\{\frac{x}{t} \in (-\infty,\lambda_1^-)\}$, $\{\frac{x}{t} \in (\lambda_l^-,\lambda_l^+)\}$ for all $l \in \{1{:}m\}$, $\{\frac{x}{t} \in (\lambda_{l-1}^+,\lambda_l^-)\}$ for all $l \in \{2{:}m\}$, and $\{\frac{x}{t} \in (\lambda_m^+,\infty)\}$, such that the Riemann solution is $\boldsymbol{u}_L$ in the first sector $\{\frac{x}{t} \in (-\infty,\lambda_1^-)\}$ and $\boldsymbol{u}_R$ in the last sector $\{\frac{x}{t} \in (\lambda_m^+,\infty)\}$, and the solution in the other sectors is either a constant state or an expansion wave. If $\lambda_l^- = \lambda_l^+$, then the corresponding l-th wave is a shock or a contact discontinuity. For all $l \in \{1{:}m\}$, the solution associated with the pair $(\lambda_l^-,\lambda_l^+)$ is called l-th wave.

Definition 80.19 (Riemann fan). *Let $(\boldsymbol{u}_L,\boldsymbol{u}_R) \in \mathcal{A}^2$ and $\boldsymbol{n} \in \mathbb{R}^d$ be a unit vector. Let $\{\lambda_n^\pm\}_{n\in\{1:m\}}$ satisfy* (80.20)*. The sector $\{\lambda_1^- < \frac{x}{t} < \lambda_m^+\}$ is henceforth called* Riemann fan*; see Figure* 80.1*.*

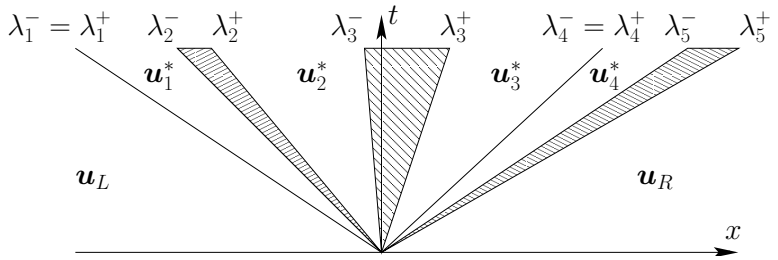

Figure 80.1: Example of a Riemann fan with $m := 5$. The 1-wave and the 4-wave are shocks or contact discontinuities, the 2-wave, the 3-wave, and the 5-wave are expansions. The states $\boldsymbol{u}_L$, $\boldsymbol{u}_1^*$, $\boldsymbol{u}_2^*$, $\boldsymbol{u}_3^*$, $\boldsymbol{u}_4^*$, $\boldsymbol{u}_R$ are constant.

Remark 80.20 (Literature). Theorem 80.18 has been first proved in Lax [212, Thm. 9.1]. A comprehensive treatment of this problem has been done in Bressan [50, Thm. 5.3]. We also refer the reader to Holden and Risebro [184, Thm. 5.17], Godlewski and Raviart [138, Thm. 6.1] for detailed proofs of this result. The case of strictly hyperbolic systems that may have eigenpairs that are neither genuinely nonlinear nor linearly degenerate is treated in, e.g., Liu [221, Thm. 1.2], Dafermos [97, Thm. 9.5.1]. In the case of general hyperbolic systems, we refer to Bianchini and Bressan [32, §14] for characterizations of the Riemann solution using viscosity regularization. We also refer to Young [290, Thm. 2] for the theory of the Riemann problem for the p-system with arbitrary data (i.e., with possible formation of vacuum) and to Toro [277, Chap. 4] for the theory of the Riemann problem for the Euler equations and a review of associated numerical methods. □

80.2.2 Maximum speed and averages

The goal of this section is to collect key notions and results on the Riemann fan that will be needed in the next chapters, where numerical approximation schemes will be constructed. First, we will need to have at our disposal a real number $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ such that

$$\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R) \ge \max(|\lambda_1^-(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)|,|\lambda_m^+(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)|), \tag{80.21}$$

where λ_1^- and λ_m^+ satisfy (80.20), i.e., these two real numbers depend on $(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ and are used to define the Riemann fan (see Definition 80.19).

Definition 80.21 (Maximum wave speed). *Let $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{A}^2$ and $\boldsymbol{n} \in \mathbb{R}^d$ be a unit vector. The number* $\max(|\lambda_1^-(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)|, |\lambda_m^+(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)|)$ *is called* maximum wave speed *in the Riemann problem. Any real number $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ satisfying* (80.21) *is called upper bound on the maximum wave speed in the Riemann problem.*

Denoting by $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ the vanishing-viscosity solution to the Riemann problem (80.13) constructed in Theorem 80.18, the first key result that we are going to use repeatedly is that this solution satisfies for all $t \geq 0$,

$$\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t) = \begin{cases} \boldsymbol{u}_L & \text{if } x \leq -t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R), \\ \boldsymbol{u}_R & \text{if } x \geq t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R). \end{cases} \tag{80.22}$$

Moreover, a quantity that will be of interest to us is the *Riemann average*

$$\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \coloneqq \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)\,\mathrm{d}x, \tag{80.23}$$

where we take $0 \leq t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$. To state the vector-valued counterpart of Lemma 79.18, we observe that if $(\eta, \boldsymbol{q})$ is an entropy pair for (80.2), then $(\eta, \boldsymbol{q}\cdot\boldsymbol{n})$ is an entropy pair for (80.13) since $\partial_{v_j}(\boldsymbol{q}(\boldsymbol{v})\cdot\boldsymbol{n}) = \sum_{i\in\{1:m\}} \partial_{v_i}\eta(\boldsymbol{v})\partial_{v_j}(\mathbb{f}(\boldsymbol{v})\cdot\boldsymbol{n})_i$ for all $\boldsymbol{v} \in \mathcal{A}$, all $j \in \{1{:}m\}$, and every unit vector $\boldsymbol{n} \in \mathbb{R}^d$.

Lemma 80.22 (Riemann average). *Recall that $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ is defined in* (80.23) *for $0 \leq t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$. Let $(\eta, \boldsymbol{q})$ be an entropy pair for* (80.2). *Then we have*

$$\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) = \frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R) - t\big(\mathbb{f}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \mathbb{f}(\boldsymbol{u}_L)\cdot\boldsymbol{n}\big), \tag{80.24a}$$

$$\eta(\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)) \leq \frac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) - t\big(\boldsymbol{q}(\boldsymbol{u}_R)\cdot\boldsymbol{n} - \boldsymbol{q}(\boldsymbol{u}_L)\cdot\boldsymbol{n}\big). \tag{80.24b}$$

Proof. To prove (80.24a), we integrate (80.13) over $(-\frac{1}{2}, \frac{1}{2})\times(0,t)$ and use that $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t) = \boldsymbol{u}_L$ if $x \leq t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ and $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t) = \boldsymbol{u}_R$ if $x \geq t\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. To prove (80.24b), we integrate (80.5) over the same set and use Jensen's inequality. □

80.2.3 Invariant sets

The notion of maximum principle is not valid in general for hyperbolic systems, even in the linear case. We refer the reader to Exercises 80.6 and 80.7 for counterexamples with the linear wave equation. Following Chueh et al. [88], Hoff [183], Smoller [263], Frid [130], we extend the notion of maximum principle to hyperbolic systems by introducing the notion of invariant set.

Definition 80.23 (Invariant set). *A convex set $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ is said to be* invariant *for the hyperbolic system* (80.2) *if for every pair $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{B}^2$ and every unit vector $\boldsymbol{n} \in \mathbb{R}^d$, the vanishing-viscosity solution to the Riemann problem* (80.13)*, $\boldsymbol{u}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)(x,t)$, is in $\overline{\mathcal{B}}$ for a.e. $x \in \mathbb{R}$ and a.e. $t > 0$ with $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$, where $\overline{\mathcal{B}}$ is the closure of $\mathcal{B}$.*

Lemma 80.24 (Riemann average). *Let $\mathcal{B} \subset \mathcal{A} \subset \mathbb{R}^m$ be an invariant set for* (80.2). *Let $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathcal{B}^2$ and $\boldsymbol{n} \in \mathbb{R}^d$ be a unit vector.* (i) *If $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \leq \frac{1}{2}$, then $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \in \overline{\mathcal{B}}$.* (ii) *If $t\,\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) < \frac{1}{2}$ and $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathrm{int}(\mathcal{B})^2$, then $\overline{\boldsymbol{u}}(t, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathrm{int}\,(\mathcal{B})$, where $\mathrm{int}(\mathcal{B})$ is the interior of $\mathcal{B}$.*

Proof. (i) The function $\mathsf{d}(\boldsymbol{v}) := \inf_{\boldsymbol{z}\in\mathcal{B}} \|\boldsymbol{v}-\boldsymbol{z}\|_{\ell^2}$ is convex since $\mathcal{B}$ is convex. Jensen's inequality gives $\mathsf{d}(\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)) \le \int_{-\frac12}^{\frac12} \mathsf{d}(\boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t))\,\mathrm{d}x = 0$ because $\mathcal{B}$ is an invariant set. This proves (i).
(ii) Let $\boldsymbol{w}(t) := \frac{1}{2\lambda_{\max}t}\int_{-\lambda_{\max}t}^{\lambda_{\max}t} \boldsymbol{u}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)(x,t)\,\mathrm{d}x$. Then we have

$$\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R) = (1-2\lambda_{\max}t)\tfrac12(\boldsymbol{u}_L+\boldsymbol{u}_R) + 2\lambda_{\max}t\boldsymbol{w}(t).$$

The same argument as above shows that $\boldsymbol{w}(t)\in\overline{\mathcal{B}}$. Since $\frac12(\boldsymbol{u}_L+\boldsymbol{u}_R)\in\mathrm{int}(\mathcal{B})$, $\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ cannot belong to $\partial\mathcal{B}$. Hence, $\overline{\boldsymbol{u}}(t,\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)\in\mathrm{int}(\mathcal{B})$. □

Example 80.25 (Scalar case). Assume that $m=1$ and d is arbitrary, i.e., (80.2) is a scalar conservation equation. Any interval $[a,b]\subset\mathbb{R}$ is an admissible set of states and is an invariant set, i.e., if $u_R,u_L\in[a,b]$, then $a\le u(\boldsymbol{n},u_L,u_R)(x,t)\le b$ for a.e. x and a.e. $t>0$. This property is called *maximum principle*; see Theorem 79.11. □

Example 80.26 (p-system). Let $w_1(\boldsymbol{u}) := u+\int_v^\infty\sqrt{-p'(s)}\,\mathrm{d}s$, $w_2(\boldsymbol{u}) := u-\int_v^\infty\sqrt{-p'(s)}\,\mathrm{d}s$ if $1<\gamma$, and $w_1(\boldsymbol{u}) := u-\sqrt{r}\ln(v)$, $w_2(\boldsymbol{u}) := u+\sqrt{r}\ln(v)$ if $1=\gamma$ (recall that $p(v) := rv^{-\gamma}$ with $r>0$ and $\gamma\ge1$; see Example 80.9). Let $a,b\in\mathbb{R}$. It can be shown that any set of the form $\mathcal{B} := \{\boldsymbol{u}\in\mathbb{R}_+\times\mathbb{R} \mid a\le w_2(\boldsymbol{u}),\ w_1(\boldsymbol{u})\le b\}$ is an invariant set for the system (80.8) for $\gamma\ge1$; see Hoff [183, Ex. 3.5, p. 597] for a proof in the context of viscous regularization and Young [290] for a direct proof. □

Example 80.27 (Euler). The set $\mathcal{B} := \{(\rho,\boldsymbol{m},E) \mid \rho>0,\ E/\rho-\frac12\boldsymbol{m}^2/\rho^2\ge0\}$ is an invariant set for the compressible Euler equations. It is shown in Exercise 80.5 that $\mathcal{B}$ is convex. Since the specific entropy satisfies $\partial_t s+\boldsymbol{u}\cdot\nabla s\ge0$, there is a minimum principle on the specific entropy, so that the set $\mathcal{B}_r := \{\boldsymbol{u}=(\rho,\boldsymbol{m}^\mathsf{T},E)^\mathsf{T} \mid \rho>0,\ e(\boldsymbol{u})\ge0,\ s(\rho,e(\boldsymbol{u}))\ge r\}$ is an invariant set for all $r\in\mathbb{R}$. It is also shown in Exercise 80.5 that $\mathcal{B}_r$ is convex. Note finally that it may be important in some situations to distinguish $\overline{\mathcal{B}}$ and $\mathrm{int}\,(\mathcal{B})$. In particular, the vacuum state $\{\rho=0\}$ and the zero energy state $\{e(\boldsymbol{u})=E/\rho-\frac12\boldsymbol{m}^2/\rho^2=0\}$ do not belong to $\mathrm{int}\,(\mathcal{B})$. □

Exercises

Exercise 80.1 (1D linear system). (i) Let $u_0\in L^\infty_{\mathrm{loc}}(\mathbb{R})$. Show that $u(x,t) := u_0(x-\lambda t)$ is a weak solution to the problem $\partial_t u+\lambda\partial_x u=0$, $u(x,0)=u_0(x)$, i.e., $\int_0^\infty\int_{\mathbb{R}} u(\partial_t\phi+\lambda\partial_x\phi)\,\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\phi(x,0)\,\mathrm{d}x = 0$ for all $\phi\in C^1_0(\mathbb{R}\times\mathbb{R}_+)$. (ii) Let $\boldsymbol{u}_0\in L^\infty_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^m)$. Consider the one-dimensional linear system $\partial_t\boldsymbol{u}+\mathbb{A}\partial_x\boldsymbol{u}=0$, $\boldsymbol{u}(x,0)=\boldsymbol{u}_0(x)$, $(x,t)\in\mathbb{R}\times\mathbb{R}_+$, where $\mathbb{A}\in\mathbb{R}^{m\times m}$ is diagonalizable in $\mathbb{R}$. Give a weak solution to this problem. (iii) Solve the 1D linear wave equation, i.e., consider $\mathbb{A} := \left(\begin{smallmatrix}0&1\\c^2&0\end{smallmatrix}\right)$.

Exercise 80.2 (Linear wave equation). Consider the matrix $\mathbb{A}(\boldsymbol{n}) := \left(\begin{smallmatrix}0&\boldsymbol{n}^\mathsf{T}\\c^2\boldsymbol{n}&0\end{smallmatrix}\right)$, where $\boldsymbol{n}$ is a unit (column) vector in $\mathbb{R}^d$. Let $\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_{d-1}\}$ be such that $\{\boldsymbol{n},\boldsymbol{v}_1,\dots,\boldsymbol{v}_{d-1}\}$ is an orthonormal basis of $\mathbb{R}^d$. Show that $(c,(1,c\boldsymbol{n})^\mathsf{T})$, $(-c,(1,-c\boldsymbol{n})^\mathsf{T})$, $(0,(0,\boldsymbol{v}_1))$,…, $(0,(0,\boldsymbol{v}_{d-1}))$ are eigenpairs of $\mathbb{A}(\boldsymbol{n})$.

Exercise 80.3 (Entropy inequality). Let $\boldsymbol{u}_\epsilon$ be the smooth function satisfying $\partial_t\boldsymbol{u}_\epsilon+\nabla\cdot\mathbb{f}(\boldsymbol{u}_\epsilon)-\epsilon\Delta\boldsymbol{u}_\epsilon=0$ in $D\times\mathbb{R}_+$, $\boldsymbol{u}_\epsilon(\cdot,0)=\boldsymbol{u}_0$ in D, with $\epsilon>0$. Let $(\eta,\boldsymbol{q})$ be an entropy pair with $\eta\in C^2(\mathbb{R}^m;\mathbb{R})$. Prove that $\partial_t\eta(\boldsymbol{u}_\epsilon)+\nabla\cdot\boldsymbol{q}(\boldsymbol{u}_\epsilon)-\epsilon\Delta\eta(\boldsymbol{u}_\epsilon)\le0$.

Exercise 80.4 (Convexity). Let $\sigma : \mathcal{T}\times\mathcal{E} \subset \mathbb{R}^2 \to \mathcal{S} \subset \mathbb{R}$ be a function of class C^2 such that $\partial_e\sigma(\tau, e) > 0$ for all $(\tau, e) \in \mathcal{T}\times\mathcal{E}$. (i) Show that there exists a function $\epsilon : \mathcal{T}\times\mathcal{S} \to \mathcal{E}$ such that $\sigma(\tau, \epsilon(\tau, s)) = s$ for all $(\tau, s) \in \mathcal{T}\times\mathcal{S}$ and ϵ is of class C^2. (ii) Show that $\epsilon(\tau, \sigma(\tau, e)) = e$ for all $(\tau, e) \in \mathcal{T}\times\mathcal{E}$. (iii) Show that the following statements are equivalent: (a) The function $\epsilon : \mathcal{T}\times\mathcal{S} \to \mathcal{E}$ is strictly convex; (b) The function $-\sigma : \mathcal{T}\times\mathcal{E} \to \mathcal{S}$ is strictly convex. (*Hint*: recall that a function $\phi : X \subset \mathbb{R}^m \to \mathbb{R}$ of class C^2 is convex in the open set X iff $D^2\phi(x)(h, h) > 0$ for all $h \in \mathbb{R}^m\backslash\{0\}$ and all $x \in X$.)

Exercise 80.5 (Euler). Recall from Example 80.10 the conserved variable $\boldsymbol{u} := (\rho, \boldsymbol{m}^\mathsf{T}, E)^\mathsf{T}$, the specific internal energy $e(\boldsymbol{u}) := E/\rho - \frac{1}{2}\boldsymbol{m}^2/\rho^2$, and the function $\Phi(\boldsymbol{u}) := s(\rho, e(\boldsymbol{u}))$, where s is the specific entropy. (i) Is the function $\boldsymbol{u} \mapsto e(\boldsymbol{u})$ convex? (ii) Set $\Psi(\boldsymbol{u}) := -\rho\Phi(\boldsymbol{u})$. It is shown in Harten et al. [180, §3] that $\rho^{-1}K(D^2\Psi)K^\mathsf{T} = -C$, where $D^2\Psi$ is the Hessian matrix of Ψ and

$$K := \begin{pmatrix} 1 & \boldsymbol{v}^\mathsf{T} & \frac{1}{2}\boldsymbol{v}^2 + e \\ \boldsymbol{0} & \rho\mathbb{I}_d & \boldsymbol{m} \\ 0 & \boldsymbol{0}^\mathsf{T} & \rho \end{pmatrix}, \qquad C := \begin{pmatrix} \partial_{\rho\rho}s + \frac{2}{\rho}\partial_\rho s & \boldsymbol{0}^\mathsf{T} & \partial_{\rho e}s \\ \boldsymbol{0} & -\partial_e s\mathbb{I}_d & \boldsymbol{0} \\ \partial_{\rho e}s & \boldsymbol{0}^\mathsf{T} & \partial_{ee}s \end{pmatrix}.$$

Verify that K is invertible and C is negative definite. Show that the function $\boldsymbol{u} \mapsto \Psi(\boldsymbol{u})$ is strictly convex. (iii) Show that the set $B := \{\boldsymbol{u} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0\}$ is convex and that the set $B_r = \{\boldsymbol{u} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0,\ \Phi(\boldsymbol{u}) \geq r\}$ is convex for all $r \in \mathbb{R}$. (See also Exercise 83.3.) (iv) Let p be the pressure. Show that $\partial_\rho p(\rho, s) > 0$. (*Hint*: see Exercise 80.4 and recall that $\mathrm{d}\epsilon = T\,\mathrm{d}s - p\,\mathrm{d}\tau$.)

Exercise 80.6 (Wave equation blowup). Consider the linear wave equation in dimension three, $\partial_t u + \nabla\cdot\boldsymbol{v} = 0$, $\partial_t\boldsymbol{v} + \nabla u = \boldsymbol{0}$, with $u(\boldsymbol{x}, 0) = u_0(\|\boldsymbol{x}\|_{\ell^2})$, $\boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{0}$. Assume that $u_0 \in C^2(\mathbb{R}_+; \mathbb{R})$. (i) Show that u must solve $\partial_{tt}u - \nabla\cdot\nabla u = 0$. (ii) Let $f : \mathbb{R} \to \mathbb{R}$ be such that $f(s) := \frac{s}{2}u_0(s)$ if $s \geq 0$ and $f(s) := -f(-s)$ if $s \leq 0$. Let us write $r := \|\boldsymbol{x}\|_{\ell^2}$ and $\boldsymbol{e}_r := \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_{\ell^2}}$ if $\boldsymbol{x} \neq \boldsymbol{0}$. Show that $u(\boldsymbol{x}, t) = \frac{f(r+t)}{r} + \frac{f(r-t)}{r}$ and $\boldsymbol{v}(\boldsymbol{x}, t) = v(r, t)\boldsymbol{e}_r$, where the function $v(r, t) := -\frac{1}{r^2}\int_0^t \big(rf'(r+\tau) - f(r+\tau) + rf'(r-\tau) - f(r-\tau)\big)\,\mathrm{d}\tau$ solves the linear wave equation. (*Hint*: use spherical coordinates.) (iii) Compute $u(0, t)$ for $t > 0$. (iv) Let $\alpha \in (\frac{1}{2}, 1)$. Let $u_0(r) := 0$ if $0 \leq r \leq 1$, $u_0(r) := (r-1)^\alpha(2-r)^2$ if $r \in [1, 2]$, and $u_0(r) := 0$ if $2 \leq r$. Show that $u(\cdot, 1)$ is unbounded but $u(\cdot, 1) \in H^1(\mathbb{R}^3)$.

Exercise 80.7 (1D linear wave equation). Consider the 1D linear wave equation $\partial_t\boldsymbol{u} + \partial_x\mathbb{f}(\boldsymbol{u}) = 0$, where $\boldsymbol{u} := (\rho, v)^\mathsf{T}$, $\mathbb{f}(\boldsymbol{u}) := (\rho_0 v, p(\rho))^\mathsf{T}$, $p(\rho) := \frac{a^2}{\rho_0}\rho$, with the constants $\rho_0 > 0$ and $a > 0$. The purpose of the exercise is to show that the maximum principle does not hold true on ρ for the linear wave equation. (i) Show that the system is strictly hyperbolic. (ii) Are the characteristic families genuinely nonlinear or linearly degenerate? (iii) Consider the Riemann problem with $\boldsymbol{u}_L := (\rho_L, v_L)^\mathsf{T}$ and $\boldsymbol{u}_R := (\rho_R, v_R)^\mathsf{T}$. Express the two eigenvectors in terms of $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$. (iv) Solve the Riemann problem. (*Hint*: the solution is composed of three constant states separated by two contact discontinuities; apply the Rankine–Hugoniot condition two times.) (v) Give a condition on $v_L - v_R$ and $\rho_L - \rho_R$ so that $\min_{x\in\mathbb{R}}\rho(x, t) < \min(\rho_L, \rho_R)$. Give a condition on $v_L - v_R$ and $\rho_L - \rho_R$ so that $\min_{x\in\mathbb{R}}\rho(x, t) > \max(\rho_L, \rho_R)$. *Note*: this exercise shows that in general the maximum principle does not hold true on ρ for the linear wave equation.

Chapter 81

First-order approximation

This chapter focuses on the approximation of nonlinear hyperbolic systems using finite elements. We describe a somewhat loose adaptation to finite elements of a scheme introduced by Lax [211, p. 163]. The method, introduced in Guermond and Nazarov [153], Guermond and Popov [157], can be informally shown to be first-order accurate in time and space and to preserve every invariant set of the hyperbolic system. The time discretization is based on the forward Euler method and the space discretization employs finite elements. The theory applies regardless of whether H^1-conforming or discontinuous elements are used. Higher-order extensions are presented in Chapter 82 and Chapter 83. We draw the attention of the reader to the fact that from now on the notation regarding time-stepping is slightly different from that used in Chapters 67 to 78. The current time step is now denoted by t_n (instead of t_{n-1}) and the update is done at t_{n+1} (instead of t_n). This choice is purely aesthetic. Since we are working with explicit methods, it is shorter to refer to current quantities with the index n than with the index $(n-1)$.

81.1 Scalar conservation equations

Although the method that we propose is the same whether the problem is a scalar conservation equation or a hyperbolic system, we start by considering scalar conservation equations for simplicity. Thus, this section is devoted to the space and time approximation of the nonlinear scalar conservation equation (79.1) posed in $D \times (0,T)$ with a domain $D \subset \mathbb{R}^d$ and $T > 0$. To simplify questions regarding boundary conditions, we assume that Dirichlet boundary conditions can be enforced in the form $u(\boldsymbol{x},t) = u_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial D$ and all $t \in [0,T]$. This is the case for instance as in §79.1.1 if there is a compact subset $S \subsetneq D$ such that $u_{0|D\setminus S}$ is constant over each connected component of $D\setminus S$ (there is only one connected component if $d \geq 2$), and there exists some time $T > 0$ such that $u(\boldsymbol{x},t) = u_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial D$ and all $t \in [0,T]$.

81.1.1 The finite element space

We want to approximate the solution to (79.1) by using finite elements in space and the forward Euler scheme in time. We first present the method with continuous finite elements for simplicity (dG extensions are discussed in Remark 81.7). Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular family of matching meshes so that each mesh covers D exactly. The meshes may be nonaffine. Let $(\widehat{K},\widehat{P},\widehat{\Sigma})$ be the reference element and let $\boldsymbol{T}_K : \widehat{K} \to K$ be the geometric mapping for all $K \in \mathcal{T}_h$. We consider the

scalar-valued finite element space (see Chapter 19)

$$P_k^{\mathrm{g}}(\mathcal{T}_h) = \{v \in C^0(\overline{D};\mathbb{R}) \mid v_{|K} \circ \boldsymbol{T}_K \in \widehat{P},\ \forall K \in \mathcal{T}_h\}. \tag{81.1}$$

The reference shape functions are denoted by $\{\widehat{\theta}_i\}_{i\in\mathcal{N}}$, with $\mathcal{N} := \{1{:}n_{\mathrm{sh}}\}$. These functions form a basis of $\widehat{P}$ with the partition of unity property $\sum_{i\in\mathcal{N}} \widehat{\theta}_i(\widehat{\boldsymbol{x}}) = 1$ for all $\widehat{\boldsymbol{x}} \in \widehat{K}$. The global shape functions in $P_k^{\mathrm{g}}(\mathcal{T}_h)$ are denoted by $\{\varphi_i\}_{i\in\mathcal{A}_h}$, where $\mathcal{A}_h := \{1{:}I\}$ and $I := \dim(P_k^{\mathrm{g}}(\mathcal{T}_h))$. Let $\mathtt{j\_dof} : \mathcal{T}_h \times \mathcal{N} \to \mathcal{A}_h$ be the connectivity array s.t. $\varphi_{\mathtt{j\_dof}(K,i)|K} = \widehat{\theta}_i \circ \boldsymbol{T}_K^{-1}$ for all $(K,i) \in \mathcal{T}_h \times \mathcal{N}$. This identity together with the partition of unity property implies that

$$\sum_{i\in\mathcal{A}_h} \varphi_i(\boldsymbol{x}) = 1, \quad \forall \boldsymbol{x} \in \overline{D}. \tag{81.2}$$

We recall that $\mathcal{A}_h$ can be partitioned as $\mathcal{A}_h = \mathcal{A}_h^\circ \cup \mathcal{A}_h^\partial$, where $\mathcal{A}_h^\circ$ is the collection of the interior nodes s.t. $\varphi_{|\partial D}$ vanishes identically. This decomposition will be invoked to handle the boundary conditions.

For all $i \in \mathcal{A}_h$, we denote

$$\mathcal{I}(i) := \{j \in \mathcal{A}_h \mid \varphi_j\varphi_i \not\equiv 0\}. \tag{81.3}$$

We observe that $j \in \mathcal{I}(i)$ iff $i \in \mathcal{I}(j)$. Let $\mathcal{M}$ be the consistent mass matrix with entries m_{ij}, and let $\overline{\mathcal{M}}$ be the diagonal lumped mass matrix with entries m_i, where for all $i \in \mathcal{A}_h$ and all $j \in \mathcal{I}(i)$,

$$m_{ij} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\,\mathrm{d}x, \qquad m_i := \int_D \varphi_i(\boldsymbol{x})\,\mathrm{d}x. \tag{81.4}$$

The partition of unity property implies that $m_i = \sum_{j\in\mathcal{I}(i)} m_{ij}$. One key assumption that we shall invoke in the rest of the chapter is that

$$m_i > 0, \qquad \forall i \in \mathcal{A}_h. \tag{81.5}$$

This property holds true for linear Lagrange elements on simplices, quadrangles and hexahedra, and for Bernstein–Bezier finite elements of any polynomial degree; see, e.g., Lai and Schumaker [210, Chap. 2], Ainsworth [5].

81.1.2 The scheme

Let $u_h^0 := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^0\varphi_i \in P_k^{\mathrm{g}}(\mathcal{T}_h)$ be a reasonable approximation of u_0 (we shall be more precise in the following sections). Let $t_n \geq 0$ be the current time, and assume we are given some time step $\tau_n > 0$ for all $n \in \mathbb{N}$. The time step may depend on n, i.e., it may vary at each time t_n, but for simplicity, we are going to write τ instead of τ_n. We also write $t_{n+1} := t_n + \tau$. The space approximation of u at time t_n for all $n \in \mathbb{N}$ is written

$$u_h^n := \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^n\varphi_i \in P_k^{\mathrm{g}}(\mathcal{T}_h). \tag{81.6}$$

Notice that our assumption on the boundary conditions means that $\mathsf{U}_i^n = \mathsf{U}_i^0$ for all $i \in \mathcal{A}_h^\partial$.

The forward Euler scheme in time consists of computing u_h^{n+1}, $n \geq 0$, once u_h^n is known. We approximate $\boldsymbol{f}(u_h^n)$ by $\sum_{j\in\mathcal{A}_h} \boldsymbol{f}(\mathsf{U}_j^n)\varphi_j$. This ansatz is exact if $\boldsymbol{f}$ is linear. Hence, the truncation error, and hopefully the approximation error, is at least informally second-order accurate in space. (Recall that the truncation error is the residual that is obtained by inserting the solution to (79.1) into the scheme.) If $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ is a Lagrange finite element of degree k and $\boldsymbol{f}$ is

smooth, the expected order of the truncation error in space is $\mathcal{O}(h^{k+1})$ (on uniform meshes) since in this case $\sum_{j\in\mathcal{A}_h}\boldsymbol{f}(\mathsf{U}_j^n)\varphi_j$ is just the Lagrange interpolant of $\boldsymbol{f}(u_h^n)$. In conclusion, we have $\int_D\nabla\cdot(\boldsymbol{f}(u_h^n))\varphi_i\,\mathrm{d}x\approx\sum_{j\in\mathcal{I}(i)}\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{c}_{ij}$, where the vector $\boldsymbol{c}_{ij}\in\mathbb{R}^d$ is defined by

$$\boldsymbol{c}_{ij}:=\int_D\varphi_i\nabla\varphi_j\,\mathrm{d}x,\qquad\forall i,j\in\mathcal{A}_h.\tag{81.7}$$

This vector is zero if $j\notin\mathcal{I}(i)$. Note that $\boldsymbol{c}_{ij}$ scales like $m_i^{1-\frac{1}{d}}$, $\boldsymbol{c}_{ii}=\boldsymbol{0}$ for all $i\in\mathcal{A}_h^\circ$, and $\boldsymbol{c}_{ij}=-\boldsymbol{c}_{ji}$ if i or j is in $\mathcal{A}_h^\circ$. Moreover, the partition of unity property implies that

$$\sum_{j\in\mathcal{I}(i)}\boldsymbol{c}_{ij}=\boldsymbol{0},\qquad\forall i\in\mathcal{A}_h.\tag{81.8}$$

Given $u_h^n\in P_k^{\mathrm{g}}(\mathcal{T}_h)$, we then compute $u_h^{n+1}:=\sum_{i\in\mathcal{A}_h}\mathsf{U}_i^{n+1}\varphi_i$ from

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau}+\sum_{j\in\mathcal{I}(i)}\Big(\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{c}_{ij}-d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)\Big)=0,\tag{81.9}$$

for all $i\in\mathcal{A}_h^\circ$, and $\mathsf{U}_i^{n+1}=\mathsf{U}_i^0$ for all $i\in\mathcal{A}_h^\partial$. The real number d_{ij}^n depends on U_i^n and U_j^n as follows: For all $i,j\in\mathcal{A}_h$ with $i\neq j$,

$$d_{ij}^n:=\max(\lambda_{\max}(\boldsymbol{n}_{ij},\mathsf{U}_i^n,\mathsf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2},\lambda_{\max}(\boldsymbol{n}_{ji},\mathsf{U}_j^n,\mathsf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}),\tag{81.10}$$

where $\boldsymbol{n}_{ij}:=\boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$ and $\lambda_{\max}(\boldsymbol{n},u_L,u_R)$ is any upper bound on the maximum wave speed in the Riemann problem with the data (u_L,u_R) and the flux $\boldsymbol{f}\cdot\boldsymbol{n}_{ij}$ as explained in §79.2. We observe that $d_{ij}^n=d_{ji}^n$ and that the definition of d_{ii}^n is irrelevant in (81.10). The coefficient d_{ij}^n is called *graph viscosity* (and sometimes also *artificial viscosity*). We prefer to employ the term "graph vicosity" since it emphasizes that d_{ij}^n is computed using the vectors $\boldsymbol{c}_{ij}$ which directly encode the mesh geometry, whereas the term "artificial viscosity" usually refers to a discrete counterpart of a viscous regularization term; see Remark 81.6 for further comments on the terminology. The actual justification for (81.10) will be given in §81.1.3 by establishing a maximum principle and in §81.2.2 by establishing an invariant domain property for hyperbolic systems.

Remark 81.1 (Mass lumping). It is important that the mass matrix be lumped in (81.9). It is indeed shown in Guermond et al. [168] that for every nonzero Lipschitz flux, there exists some initial data $\{\mathsf{U}_i^0\}_{i\in\mathcal{A}_h}$ such that $\{\mathsf{U}_i^1\}_{i\in\mathcal{A}_h}$ violates the maximum principle for every choice of d_{ij}^0 and for all $\tau>0$ when the consistent mass matrix is used. □

Remark 81.2 (Alternative formulation). Notice that the summation in (81.9) can be reduced to $j\in\mathcal{I}(i)\backslash\{i\}$ and that $\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{c}_{ij}$ can be replaced by $\big(\boldsymbol{f}(\mathsf{U}_j^n)-\boldsymbol{f}(\mathsf{U}_i^n)\big)\cdot\boldsymbol{c}_{ij}$ owing to (81.8). □

Remark 81.3 ((81.10)). The two terms on the right-hand side of (81.10) are equal if i or j is in $\mathcal{A}_h^\circ$; see Exercise 81.2. The definition (81.10) is useful to handle general boundary conditions. □

Remark 81.4 (Conservation). Notice that $\sum_{i\in\mathcal{A}_h}m_i\mathsf{U}_i^n=\int_D u_h^n\,\mathrm{d}x$ and

$$\sum_{i\in\mathcal{A}_h}\sum_{j\in\mathcal{I}(i)}d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)=\sum_{i\in\mathcal{A}_h}\sum_{j\in\mathcal{I}(i)}d_{ij}^n\mathsf{U}_j^n-\sum_{j\in\mathcal{A}_h}\sum_{i\in\mathcal{I}(j)}d_{ij}^n\mathsf{U}_i^n=0,$$

because $d_{ij}^n=d_{ji}^n$ and $j\in\mathcal{I}(i)$ iff $i\in\mathcal{I}(j)$. This implies that $\int_D u_h^{n+1}\,\mathrm{d}x=\int_D u_h^n\,\mathrm{d}x-\tau\int_D\nabla\cdot\big(\sum_{j\in\mathcal{A}_h}\boldsymbol{f}(\mathsf{U}_j^n)\varphi_j\big)\,\mathrm{d}x$, for all $n\geq0$. Since $\mathsf{U}_j^n=\mathsf{U}_j^0$ for all $j\in\mathcal{A}_h^\partial$, and assuming that $\boldsymbol{f}(\mathsf{U}_j^0)\cdot\boldsymbol{n}_j=0$ with $\boldsymbol{n}_j:=m_j^{-1}\int_{\partial D}\boldsymbol{n}\varphi_j\,\mathrm{d}s$ for all $j\in\mathcal{A}_h^\partial$, the divergence formula yields $\int_D u_h^{n+1}\,\mathrm{d}x=\int_D u_h^0\,\mathrm{d}x$. One says that the scheme (81.9) is (globally) conservative. □

Example 81.5 (1D, linear transport). Let $D := (-1,1)$ and $\boldsymbol{f}(v) := f(v)\boldsymbol{e}_x$, where $\boldsymbol{e}_x$ is the unit vector orienting $\mathbb{R}$. Let $\mathcal{T}_h$ be the mesh composed of the cells $\{[x_i,x_{i+1}]\}_{i\in\{1:I-1\}}$ with the convention $x_1 := -1$, $x_I := 1$, so that $\mathcal{A}_h = \{1{:}I\}$, $\mathcal{A}_h^\circ = \{2{:}I{-}1\}$, and $\mathcal{A}_h^\partial = \{1,I\}$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the space composed of the continuous piecewise linear functions on $\mathcal{T}_h$. Assuming that $i\in\mathcal{A}_h^\circ$, we have $\boldsymbol{c}_{ii}=\boldsymbol{0}$, $\boldsymbol{c}_{i,i-1}=-\frac{1}{2}\boldsymbol{e}_x$, $\boldsymbol{c}_{i,i+1}=\frac{1}{2}\boldsymbol{e}_x$, and $m_i=\frac{h_{i-1}+h_i}{2}$ with $h_i := x_{i+1}-x_i$. The scheme (81.9) becomes for all $i\in\mathcal{A}_h^\circ$,

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau} = \frac{f(\mathsf{U}_{i-1}^n)-f(\mathsf{U}_{i+1}^n)}{2} + d_{i,i-1}^n(\mathsf{U}_{i-1}^n-\mathsf{U}_i^n) + d_{i,i+1}^n(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n),$$

together with $\mathsf{U}_1^{n+1}=\mathsf{U}_1^n$ and $\mathsf{U}_I^{n+1}=\mathsf{U}_I^n$. Let us take $\boldsymbol{f}(u)=\beta u\boldsymbol{e}_x$. Then $\lambda_{\max}(u_L,u_R)=|\beta|$ for all $u_L,u_R\in\mathbb{R}$ and $d_{i,i-1}=\frac{1}{2}|\beta|=d_{i,i+1}$. The scheme thus reduces to the classical *upwind* approximation:

$$m_i\frac{\mathsf{U}_i^{n+1}-\mathsf{U}_i^n}{\tau} = \frac{1}{2}(\beta+|\beta|)(\mathsf{U}_{i-1}^n-\mathsf{U}_i^n) + \frac{1}{2}(|\beta|-\beta)(\mathsf{U}_{i+1}^n-\mathsf{U}_i^n). \tag{81.11}$$

For instance, $\frac{m_i}{\tau}(\mathsf{U}_i^{n+1}-\mathsf{U}_i^n)=\beta(\mathsf{U}_{i-1}^n-\mathsf{U}_i^n)$ if $\beta\ge 0$. Note finally that we also have

$$-\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n) = \frac{|\beta|h_{i-1}}{2}\int_{x_{i-1}}^{x_i}\nabla u_h^n\cdot\nabla\varphi_i\,\mathrm{d}x + \frac{|\beta|h_i}{2}\int_{x_i}^{x_{i+1}}\nabla u_h^n\cdot\nabla\varphi_i\,\mathrm{d}x.$$

Thus, $-\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)$ can be viewed as the discrete counterpart of $-\nabla\cdot(\epsilon\nabla u_\epsilon)$ with $\epsilon_{|[x_l,x_{l+1}]} := \frac{1}{2}|\beta|h_l$ for all $l\in\{1{:}I{-}1\}$. □

Remark 81.6 (Graph viscosity vs. viscous regularization). The name "artificial viscosity" given to the term $-\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)$ has its origin in the following observations. Let $\partial_t u_\epsilon + \nabla\cdot(\boldsymbol{f}(u_\epsilon)) - \epsilon\Delta u_\epsilon = 0$ be the viscous regularization of (79.1) with $\epsilon>0$. Denoting by $u_{\epsilon h}^n$ the finite element approximation of u_ϵ at t_n, the discrete counterpart of $-\epsilon\Delta u_\epsilon^n$ is $\epsilon\int_D\nabla u_{\epsilon h}^n\nabla\varphi_i\,\mathrm{d}x = \sum_{j\in\mathcal{I}(i)}\epsilon\mathsf{U}_{\epsilon j}^n\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x$. Adopting the notation $\gamma_{ij} := \epsilon\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x$ and observing that the partition of unity implies $\sum_{j\in\mathcal{I}(i)}\gamma_{ij}=0$, we have $\epsilon\int_D\nabla u_{\epsilon h}^n\nabla\varphi_i\,\mathrm{d}x = \sum_{j\in\mathcal{I}(i)}\gamma_{ij}(\mathsf{U}_j^n-\mathsf{U}_i^n)$. Referring to the material of §33.2 on the discrete maximum principle for elliptic equations, we recall that it is essential that $\gamma_{ij}=\epsilon\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x\le 0$ to satisfy the discrete maximum principle for the continuous $\mathbb{P}_1$-approximation of elliptic equations. The same phenomenon happens here: we will see in §81.1.3 that it is essential that $\gamma_{ij}\le 0$ for all $j\ne i$. Therefore, we make the change of notation $d_{ij}^n := -\gamma_{ij}$, so that $d_{ij}^n\ge 0$ for all $j\ne i$, and we have $\epsilon\int_D\nabla u_{\epsilon h}^n\nabla\varphi_i\,\mathrm{d}x = -\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\mathsf{U}_j^n-\mathsf{U}_i^n)$, which is exactly the expression used in (81.9). The analogy stops here because the definition $d_{ij}^n := -\epsilon\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x$ has two major flaws. The first one is that the condition $d_{ij}^n := -\epsilon\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x\ge 0$ requires unacceptable constraints on the mesh like the acute angle condition (see Lemma 33.9 and Definition 33.11). The second one is that we know that ϵ should go to zero but we do not a priori know how ϵ should go to zero in terms of the meshsize. Taking inspiration from Example 81.5, one could come up with some reasonable heuristics, but a better strategy consists of abandoning the definition $d_{ij}^n := -\epsilon\int_D\nabla\varphi_j\cdot\nabla\varphi_i\,\mathrm{d}x$ in favor of (81.10), since we will see in §81.1.3 that the definition (81.10) does not require any angle condition on the mesh or any ad hoc heuristics on ϵ for the maximum principle to be satisfied. □

Remark 81.7 (Extensions to dG). Notice that the only places where the finite element structure intervenes in (81.9) are the definition of the lumped mass mass coefficients m_i and the definition of the coefficients $\boldsymbol{c}_{ij}$. All that is said above can be extended to discontinuous finite elements provided the coefficients m_i and $\boldsymbol{c}_{ij}$ are defined accordingly, and $(\widehat{K},\widehat{P},\widehat{\Sigma})$ is a Lagrange element or is close to being a Lagrange element (e.g., Bernstein–Bezier finite element), as further discussed in Guermond et al. [171, §4.3]. □

81.1.3 Maximum principle

In this section, we establish an important stability property of the scheme (81.9) with d_{ij}^n defined in (81.10). For every unit vector $\boldsymbol{n} \in \mathbb{R}^d$ and every pair $(u_L, u_R) \in \mathbb{R}\times\mathbb{R}$, we denote by $\lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ any upper bound on the maximum wave speed in the one-dimensional Riemann problem with the data (u_L, u_R) and the flux $\boldsymbol{f}\cdot\boldsymbol{n}$ as explained in §79.2. We denote $\overline{u}(t, \boldsymbol{n}, u_L, u_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} u(\boldsymbol{n}, u_L, u_R)(x,t)\,\mathrm{d}x$ the Riemann average, where $u(\boldsymbol{n}, u_L, u_R)(x,t)$ is the solution to the one-dimensional Riemann problem (see Lemma 79.18).

Theorem 81.8 (Local maximum principle for components). *Let $n \in \mathbb{N}$. Assume that the entries of the lumped mass matrix are positive, i.e., that* (81.5) *holds true. Assume that τ is small enough so that the following* CFL condition *is satisfied:*

$$\min_{i\in\mathcal{A}_h^\circ}\left(1 + 2\tau\frac{d_{ii}^n}{m_i}\right) \geq 0, \tag{81.12}$$

where $d_{ii}^n := -\sum_{j\in\mathcal{I}(i)\setminus\{i\}} d_{ij}^n$ (notice that $d_{ii}^n \leq 0$). The following local maximum principle is satisfied: For all $i \in \mathcal{A}_h$,

$$\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}], \qquad \mathsf{U}_i^{\mathrm{m},n} := \min_{j\in\mathcal{I}(i)} \mathsf{U}_j^n, \qquad \mathsf{U}_i^{\mathrm{M},n} := \max_{j\in\mathcal{I}(i)} \mathsf{U}_j^n. \tag{81.13}$$

Proof. The assertion (81.13) is obviously satisfied for all $i \in \mathcal{A}_h^\partial$, so let us focus on $i \in \mathcal{A}_h^\circ$. Using that $\sum_{j\in\mathcal{I}(i)} \boldsymbol{f}(\mathsf{U}_i^n)\cdot\boldsymbol{c}_{ij} = 0$ owing to (81.8), we rewrite (81.9) as follows:

$$\mathsf{U}_i^{n+1} = \left(1 - \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^n}{m_i}\right)\mathsf{U}_i^n + \sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^n}{m_i}\overline{\mathsf{U}}_{ij}^n, \tag{81.14}$$

with $\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_i^n + \mathsf{U}_j^n) - (\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^n}$. The first key observation is that (81.14) is a convex combination if τ is small enough so that (81.12) holds true. Hence, we only need to ascertain that $\min(\mathsf{U}_i^n, \mathsf{U}_j^n) \leq \overline{\mathsf{U}}_{ij}^n \leq \max(\mathsf{U}_i^n, \mathsf{U}_j^n)$. The second key observation is that setting $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$ and introducing the fake time $t_{ij} := \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^n$, we realize that $\overline{\mathsf{U}}_{ij}^n = \overline{u}(t_{ij}, \boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)$, as established in Lemma 79.18 provided that $t_{ij}\lambda_{\max}(\boldsymbol{n}_{ij}, u_L, u_R) \leq \frac{1}{2}$. (Let us emphasize that the time t_{ij} is related to the Riemann problem (79.23) with the data $(\mathsf{U}_i^n, \mathsf{U}_j^n)$, and that this time has nothing to do with the current time t_n.) Using (81.10), we have

$$t_{ij}\lambda_{\max}(\boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n) = \frac{\|\boldsymbol{c}_{ij}\|_{\ell^2}}{2d_{ij}^n}\lambda_{\max}(\boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n) \leq \frac{1}{2}.$$

Hence, the above condition on t_{ij} is satisfied, and this implies that $\min(\mathsf{U}_i^n, \mathsf{U}_j^n) \leq \overline{\mathsf{U}}_{ij}^n \leq \max(\mathsf{U}_i^n, \mathsf{U}_j^n)$ since we have ascertained that $\overline{\mathsf{U}}_{ij}^n = \overline{u}(t_{ij}, \boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)$ and the Riemann average satisfies the maximum principle. □

The result of Theorem 81.8 holds true for the coordinate vector U^{n+1}, but we do not know yet whether this property holds true for the scalar field u_h^{n+1}. In order to infer some information on the approximate solution u_h^{n+1}, we introduce an additional assumption on the reference shape functions. More specifically, we assume that the basis $\{\widehat{\theta}_i\}_{i\in\mathcal{N}}$ is nonnegative, i.e., $\widehat{\theta}_i(\widehat{\boldsymbol{x}}) \geq 0$ for all $\widehat{\boldsymbol{x}} \in \widehat{K}$ and all $i \in \mathcal{N}$. This property holds true for linear Lagrange elements on simplices, quadrangles, hexahedra, and prisms, and for Bernstein–Bezier finite elements of any polynomial degree; see, e.g., Lai and Schumaker [210, Chap. 2], Ainsworth [5].

Corollary 81.9 (Maximum principle for discrete functions). *Let $N \in \mathbb{N}\backslash\{0\}$. Assume that $\widehat{\theta}_i(\widehat{\boldsymbol{x}}) \geq 0$ for all $\widehat{\boldsymbol{x}} \in \widehat{K}$ and all $i \in \mathcal{N}$ and that the CFL condition* (81.12) *is satisfied for all $n < N$. Let $\mathsf{U}^0_{\min} := \min_{j\in\mathcal{A}_h} \mathsf{U}^0_j$ and $\mathsf{U}^0_{\max} := \max_{j\in\mathcal{A}_h} \mathsf{U}^0_j$. Then we have*

$$u_h^n(\boldsymbol{x}) \in [\mathsf{U}^0_{\min}, \mathsf{U}^0_{\max}], \qquad \forall \boldsymbol{x} \in D,\ \forall n \in \{0{:}N\}. \tag{81.15}$$

Proof. For all $\boldsymbol{x} \in D$, we have $u_h^n(\boldsymbol{x}) = \sum_{i\in\mathcal{A}_h} \mathsf{U}_i^n \varphi_i(\boldsymbol{x})$, so that $u_h^n(\boldsymbol{x})$ is in the convex hull of $\{\mathsf{U}_i^n\}_{i\in\mathcal{A}_h}$ owing to the partition of unity property (81.2) and the nonnegativity assumption on the reference shape functions (i.e., $\varphi_i(\boldsymbol{x}) \geq 0$ for all $i \in \mathcal{A}_h$). By arguing by induction and by invoking Theorem 81.8, one deduces that the convex hull of $\{\mathsf{U}_i^n\}_{i\in\mathcal{A}_h}$ is in the convex set $[\mathsf{U}^0_{\min}, \mathsf{U}^0_{\max}]$. The assertion (81.15) follows readily. □

Remark 81.10 (Construction of u_h^0). Let $u_{\min} := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} u_0(\boldsymbol{x})$ and $u_{\max} := \operatorname{ess\,sup}_{\boldsymbol{x}\in D} u_0(\boldsymbol{x})$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be built using $\mathbb{P}_1$ Lagrange elements. Then defining u_h^0 to be the Lagrange interpolant of u_0, we have $[\mathsf{U}^0_{\min}, \mathsf{U}^0_{\max}] \subset [u_{\min}, u_{\max}]$. Similarly, if $P_k^{\mathrm{g}}(\mathcal{T}_h)$ is built using Bernstein–Bezier finite elements of degree two or higher, then defining u_h^0 to be the Bernstein–Bezier interpolant of u_0, we also have $[\mathsf{U}^0_{\min}, \mathsf{U}^0_{\max}] \subset [u_{\min}, u_{\max}]$; see [210, Eq. (2.72)]. In both cases, the discrete maximum principle (81.15) from Corollary 81.9 is satisfied. □

Remark 81.11 (Literature). A quantity similar to the *Riemann average* $\overline{\mathsf{U}}_{ij}^n$ is introduced in Lax's seminal paper [211, p. 163]. The argument invoking the convex combination (81.14) and the Riemann averages $\overline{\mathsf{U}}_{ij}^n$ can be traced back to the proof of Corollary 1 in Hoff [182]. This argument is also invoked in Harten et al. [179], Tadmor [268, p. 375], Perthame and Shu [237, Thm. 3]. The CFL condition (81.12) is named after Courant, Friedrichs, and Lewy [§II.2][91] (see also [92, §II.2, p. 228] for the English translation). □

81.1.4 Entropy inequalities

We now show that the proposed scheme satisfies discrete entropy inequalities.

Theorem 81.12 (Entropy). *Let $n \in \mathbb{N}$. Assume that the CFL condition* (81.12) *is satisfied. Let $(\eta, \boldsymbol{q})$ be an entropy pair for* (79.1). *Then the following discrete entropy inequality holds true for all $i \in \mathcal{A}_h^\circ$:*

$$\frac{m_i}{\tau}\big(\eta(\mathsf{U}_i^{n+1}) - \eta(\mathsf{U}_i^n)\big) + \int_D \nabla\cdot\Big(\sum_{j\in\mathcal{I}(i)} \boldsymbol{q}(\mathsf{U}_j^n)\varphi_j\Big)\varphi_i \,\mathrm{d}x - \sum_{j\in\mathcal{I}(i)} d_{ij}^n\big(\eta(\mathsf{U}_j^n) - \eta(\mathsf{U}_i^n)\big) \leq 0. \tag{81.16}$$

Proof. Recalling that (81.14) is a convex combination owing to the CFL condition and using the convexity of η, we infer that

$$\eta(\mathsf{U}_i^{n+1}) \leq \Big(1 - \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \frac{2\tau d_{ij}^n}{m_i}\Big)\eta(\mathsf{U}_i^n) + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \frac{2\tau d_{ij}^n}{m_i}\eta(\overline{\mathsf{U}}_{ij}^n).$$

Since $\boldsymbol{f}$ is Lipschitz, the assumptions of Lemma 79.18 hold true. Since we have shown that $\overline{\mathsf{U}}_{ij}^n = \overline{u}(t_{ij}, \boldsymbol{n}_{ij}, \mathsf{U}_i^n, \mathsf{U}_j^n)$ with the fake time $t_{ij} = \|\boldsymbol{c}_{ij}\|_{\ell^2}/2d_{ij}^n$, the inequality (79.22b) implies that

$$\eta(\overline{\mathsf{U}}_{ij}^n) \leq \frac{1}{2}(\eta(\mathsf{U}_i^n) + \eta(\mathsf{U}_j^n)) - t_{ij}(\boldsymbol{q}(\mathsf{U}_j^n)\cdot\boldsymbol{n}_{ij} - \boldsymbol{q}(\mathsf{U}_i^n)\cdot\boldsymbol{n}_{ij}).$$

Rearranging the terms leads to

$$\frac{m_i}{\tau}\big(\eta(\mathsf{U}_i^{n+1})-\eta(\mathsf{U}_i^n)\big) \le \sum_{j\in\mathcal{I}(i)\backslash\{i\}} 2d_{ij}^n\big(\eta(\overline{\mathsf{U}}_{ij}^n)-\eta(\mathsf{U}_i^n)\big)$$
$$\le \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \Big(d_{ij}^n\big(\eta(\mathsf{U}_j^n)-\eta(\mathsf{U}_i^n)\big) - \|\boldsymbol{c}_{ij}\|_{\ell^2}(\boldsymbol{q}(\mathsf{U}_j^n)\cdot\boldsymbol{n}_{ij}-\boldsymbol{q}(\mathsf{U}_i^n)\cdot\boldsymbol{n}_{ij})\Big).$$

The conclusion follows from the definitions of $\boldsymbol{n}_{ij}$ and $\boldsymbol{c}_{ij}$ and by observing that the summation can be extended to all $j\in\mathcal{I}(i)$. □

Remark 81.13 (Global bounds). Under appropriate boundary conditions implying that there is no entropy flux at the boundary ∂D, the discrete entropy inequalities (81.16) for all $i\in\mathcal{A}_h^\circ$ lead to the global entropy inequality $\sum_{i\in\mathcal{A}_h} m_i\eta(\mathsf{U}_i^{n+1}) \le \sum_{i\in\mathcal{A}_h} m_i\eta(\mathsf{U}_i^n)$ for every convex function η. □

81.2 Hyperbolic systems

In this section, we describe the time and space approximation of the hyperbolic system (80.2). To simplify the argumentation, we assume as in §81.1 that Dirichlet boundary conditions using the initial condition are enforced at the boundary. From now on, for every unit vector $\boldsymbol{n}$ in $\mathbb{R}^d$ and every pair $(\boldsymbol{u}_L,\boldsymbol{u}_R)$ in $\mathbb{R}^m\times\mathbb{R}^m$, $\lambda_{\max}(\boldsymbol{n},\boldsymbol{u}_L,\boldsymbol{u}_R)$ denotes either the maximum wave speed in the Riemann problem (80.13) or any upper bound thereof (see (80.21)). Examples of how to compute such an upper bound will be presented in §81.2.3.

81.2.1 The finite element space

The setting for the approximation in space is the same as in §81.1.1. Recall that $m\in\mathbb{N}\backslash\{0\}$ is the number of components in the hyperbolic system (80.2). Given a shape-regular family of matching meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$ so that each mesh covers D exactly, we introduce the finite element space

$$\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h) := \left(P^{\mathrm{g}}(\mathcal{T}_h)\right)^m, \tag{81.17}$$

and we still denote by $\{\varphi_i\}_{i\in\mathcal{A}_h}$ the scalar-valued basis functions of $P_k^{\mathrm{g}}(\mathcal{T}_h)$. Denoting by $(\boldsymbol{e}_k)_{k\in\{1:m\}}$ the canonical basis of $\mathbb{R}^m$, we use $\{\varphi_i\boldsymbol{e}_k\}_{i\in\mathcal{A}_h,k\in\{1:m\}}$ as a basis for $\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$. Notice that all the components in $\mathbb{R}^m$ are associated with the same scalar-valued basis function. One says that the dependent variables are *collocated*.

81.2.2 The scheme

Let $\boldsymbol{u}_h^0 := \sum_{i\in\mathcal{A}_h}\mathbf{U}_i^0\varphi_i \in \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$, with $\mathbf{U}_i^0\in\mathbb{R}^m$ for all $i\in\mathcal{A}_h$, be a reasonable approximation of $\boldsymbol{u}_0$. Let $n\in\mathbb{N}$, τ be the time step, t_n be the current time, and let us set $t_{n+1} := t_n+\tau$ (as above the time step τ may depend on n). Let $\boldsymbol{u}_h^n := \sum_{i\in\mathcal{A}_h}\mathbf{U}_i^n\varphi_i \in \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$, with $\mathbf{U}_i^n\in\mathbb{R}^m$ for all $i\in\mathcal{A}_h$, be the approximation at the discrete time t_n. Note that the coordinate vector of $\boldsymbol{u}_h^n$ is in $(\mathbb{R}^m)^I\equiv\mathbb{R}^{mI}$. We compute $\boldsymbol{u}_h^{n+1}$ by means of the forward Euler scheme as follows: For all $i\in\mathcal{A}_h^\circ$,

$$m_i\frac{\mathbf{U}_i^{n+1}-\mathbf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(i)}\Big(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - d_{ij}^n(\mathbf{U}_j^n-\mathbf{U}_i^n)\Big) = \mathbf{0}, \tag{81.18}$$

with the following *graph viscosity* coefficients:

$$d_{ij}^n := \max(\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2}, \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}), \tag{81.19}$$

for all $j \in \mathcal{I}(i)\backslash\{i\}$, recalling that $\boldsymbol{n}_{ij} := \boldsymbol{c}_{ij}/\|\boldsymbol{c}_{ij}\|_{\ell^2}$, and for all $i \in \mathcal{A}_h^\partial$, we enforce $\mathbf{U}_i^{n+1} = \mathbf{U}_i^0$.

We now generalize Theorem 81.8 and Corollary 81.9 to hyperbolic systems. Recall that $\mathcal{A} \subset \mathbb{R}^m$ denotes an admissible set for the hyperbolic system and $\mathcal{B} \subset \mathcal{A}$ an invariant set (see §80.2.3).

Theorem 81.14 (Invariant set for components). *Let $n \in \mathbb{N}$. Assume that the entries of the lumped mass matrix are positive, i.e.,* (81.5) *holds true.* (i) *Under the* CFL condition (81.12)*, i.e., $\min_{i\in\mathcal{A}_h^\circ}(1 + 2\tau\frac{d_{ii}^n}{m_i}) \geq 0$, we have*

$$[\,\{\mathbf{U}_i^n\}_{i\in\mathcal{A}_h} \subset \overline{\mathcal{B}}\,] \implies [\,\{\mathbf{U}_i^{n+1}\}_{i\in\mathcal{A}_h} \subset \overline{\mathcal{B}}\,]. \tag{81.20}$$

(ii) *Under the* tighter CFL condition $\min_{i\in\mathcal{A}_h^\circ}(1 + 2\tau\frac{d_{ii}^n}{m_i}) > 0$*, we have*

$$[\,\{\mathbf{U}_i^n\}_{i\in\mathcal{A}_h} \subset \mathrm{int}(\mathcal{B})\,] \implies [\,\{\mathbf{U}_i^{n+1}\}_{i\in\mathcal{A}_h} \subset \mathrm{int}(\mathcal{B})\,]. \tag{81.21}$$

Proof. We proceed as in the proof of Theorem 81.8. The only difference is that now $\overline{\mathbf{U}}_{ij}^{n+1}$ is either in $\overline{\mathcal{B}}$ or $\mathrm{int}(\mathcal{B})$ owing to Lemma 80.24. The CFL condition implies that $\mathbf{U}_i^{n+1}$ is a convex combination of objects that are all either in $\overline{\mathcal{B}}$ or $\mathrm{int}(\mathcal{B})$. This proves the claim since $\mathcal{B}$ is a convex set. □

Corollary 81.15 (Invariant set for discrete functions). *Let $N \in \mathbb{N}\backslash\{0\}$. Assume that the reference shape functions satisfy $\widehat{\theta}_i \geq 0$ for all $i \in \mathcal{N}$.* (i) *Assume the CFL condition* (81.12) *for all $n < N$ and that $\{\mathbf{U}_i^0\}_{i\in\mathcal{A}_h} \subset \overline{\mathcal{B}}$. Then $\{\mathbf{U}_i^n\}_{i\in\mathcal{A}_h} \subset \overline{\mathcal{B}}$ and $\boldsymbol{u}_h^n$ takes values in $\overline{B}$ for all $n \in \{0{:}N\}$.* (ii) *Assume the tighter CFL condition $\min_{i\in\mathcal{A}_h^\circ}(1 + 2\tau\frac{d_{ii}^n}{m_i}) > 0$ for all $n < N$ and that $\{\mathbf{U}_i^0\}_{i\in\mathcal{A}_h} \in \mathrm{int}(\mathcal{B})$. Then $\{\mathbf{U}_i^n\}_{i\in\mathcal{A}_h} \subset \mathrm{int}(\mathcal{B})$ and*

$$\boldsymbol{u}_h^n(\boldsymbol{x}) \in \mathrm{int}(\mathcal{B}), \qquad \forall \boldsymbol{x} \in D,\ \forall n \in \{0{:}N\}. \tag{81.22}$$

Proof. Similar to that of Corollary 81.9. □

Remark 81.16 ($\overline{\mathcal{B}}$ vs. int($\mathcal{B}$)). The distinction between $\overline{\mathcal{B}}$ and $\mathrm{int}(\mathcal{B})$ in the above statements may look a little bit pedantic, but there are applications where it is easier to work with $\mathrm{int}(\mathcal{B})$ than with $\overline{\mathcal{B}}$. For instance, for the compressible Euler equations, the invariant set $\mathcal{B}$ defined in Example 80.27 allows the vacuum state $\{\rho = 0\}$ and the state $\{e = 0\}$ in $\overline{\mathcal{B}}$, whereas $\mathrm{int}(\mathcal{B})$ does not. Although theoretically admissible, these two states may pose serious numerical difficulties. For instance, defining the velocity $\boldsymbol{u} = \boldsymbol{m}/\rho$ is problematic when $\rho \downarrow 0$. The same type of difficulty arises when estimating the specific entropy of a polytropic ideal gas, $s(\rho, e) = \ln(e^{\frac{1}{\gamma-1}}\rho^{-1})$, for $\rho \downarrow 0$ and/or $e \downarrow 0$. In conclusion, if the initial state $\{\mathbf{U}_i^0 \mid i \in \mathcal{A}_h\}$ does not contain the states $\{\rho = 0\}$ and $\{e = 0\}$, the scheme (81.18) together with the graph viscosity coefficients (81.19) and the above CFL condition never produces the states $\{\rho = 0\}$ and $\{e = 0\}$. □

81.2.3 Upper bounds on $\lambda_{\max}$

To make the algorithm (81.18)-(81.19) fully computable, it is important to have guaranteed upper bounds on $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. By going through examples, we show in this section that it is not necessary to solve the Riemann problem exactly to derive a guaranteed upper bound on $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$.

Example 81.17 (p-system). Let $\boldsymbol{u}_L := (v_L, u_L), \boldsymbol{u}_R := (v_R, u_R) \in \mathbb{R}_+ \times \mathbb{R}$ be some Riemann data for the p-system; see Example 80.9. According to Theorem 80.18, the solution to the Riemann problem consists of three constant states $\boldsymbol{u}_L$, $\boldsymbol{u}^* := (v^*, u^*)$, and $\boldsymbol{u}_R$ connected by two waves. The first wave connects $\boldsymbol{u}_L$ with $\boldsymbol{u}^*$, and the second wave connects $\boldsymbol{u}^*$ with $\boldsymbol{u}_R$. It can be shown (see, e.g., Godlewski and Raviart [138, Thm. 7.1, p. 92], Young [290], Guermond and Popov [157, Lem. 2.5]) that

$$\max(|\lambda_1^-|, |\lambda_2^+|) = \begin{cases} \sqrt{-p'(\min(v_L, v_R))} & \text{if } u_L - u_R > a, \\ \sqrt{-p'(v^*)} & \text{otherwise,} \end{cases} \tag{81.23}$$

where $a := \sqrt{(v_L - v_R)(p(v_R) - p(v_L))}$ and v^* is the unique solution of $\phi(v) := f_L(v) + f_R(v) + u_L - u_R = 0$ and f_Z, $Z \in \{L, R\}$, defined by

$$f_Z(v) := \begin{cases} -\sqrt{(p(v) - p(v_Z)(v_Z - v)} & \text{if } v \le v_Z, \\ \displaystyle\int_{v_Z}^{v} \sqrt{-p'(s)}\, \mathrm{d}s & \text{if } v > v_Z. \end{cases} \tag{81.24}$$

If $\lim_{v\to\infty} \phi(v) \le 0$, vacuum forms, and the equation $\phi(v) = 0$ has no solution. In this case, we conventionally set $v^* := \infty$ and $\sqrt{-p'(v^*)} := 0$. Solving $\phi(v^*) = 0$ can be done numerically, but an alternative to the numerical evaluation of v^* consists of estimating v^* from below as follows. Let $w_1^{\max} := \max(w_1(\boldsymbol{u}_L), w_1(\boldsymbol{u}_R))$ and $w_2^{\min} := \min(w_2(\boldsymbol{u}_L), w_2(\boldsymbol{u}_R))$, where the two functions w_1 and w_2 are defined in Example 80.26. Then let $\widetilde{\boldsymbol{u}}^* := (\widetilde{v}^*, \widetilde{u}^*)$ be the unique state such that $w_1^{\max} = w_1(\widetilde{\boldsymbol{u}}^*)$ and $w_2^{\min} = w_2(\widetilde{\boldsymbol{u}}^*)$. Assuming $\gamma > 1$, an easy computation gives

$$\widetilde{v}^* := (\gamma r)^{\frac{1}{\gamma-1}} \left(\frac{4}{(\gamma-1)(w_1^{\max} - w_2^{\min})} \right)^{\frac{2}{(\gamma-1)}}. \tag{81.25}$$

But the invariant set property guarantees that $\widetilde{v}^* \le v(\boldsymbol{u}_R, \boldsymbol{u}_L)(x,t)$ for all $x \in \mathbb{R}$ and all $t > 0$, so that $\widetilde{v}^* \le v^*$. In conclusion, replacing v^* by $\widetilde{v}^*$ in (81.23) gives the following upper bound on the maximum wave speed:

$$\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \begin{cases} \sqrt{-p'(\min(v_L, v_R))} & \text{if } u_L - u_R > a, \\ \sqrt{-p'(\widetilde{v}^*)} & \text{otherwise.} \end{cases} \tag{81.26}$$

This construction is illustrated in Figure 81.1. □

Example 81.18 (Euler equations). We refer to Guermond and Popov [156, Lem. 4.3] for an upper bound on $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ in the case of the Euler equations with the co-volume equation of state $(1 - b\rho)p = (\gamma - 1)\rho e$, with $b \ge 0$. □

Example 81.19 (Shallow-water equations). We refer to Guermond et al. [170, Lem. 4.1] for an upper bound on $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ for the shallow-water equations. □

Remark 81.20 (Average matrix). Let $\mathcal{A} \subset \mathbb{R}^m$ and let $\mathbb{f} : \mathcal{A} \to \mathbb{R}^{m\times d}$ be a Lipschitz flux with components $(\mathbb{f}_{kl})_{k\in\{1:m\}, l\in\{1:d\}}$. Let $(\mathbb{f}\cdot\boldsymbol{n})_k$ for all $k \in \{1{:}d\}$ be the components of $\mathbb{f}\cdot\boldsymbol{n}$ for every unit vector $\boldsymbol{n} \in \mathbb{R}^d$. Consider the average matrix $\mathbb{A} \in \mathbb{R}^{m\times m}$ s.t. $\mathbb{A}_{kk'} = \int_0^1 \partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))\, \mathrm{d}\theta$ (this matrix depends on the triple $(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ but for simplicity, we just write $\mathbb{A}$). By definition of hyperbolicity (see Definition 80.1), the matrix with entries $\partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k$ is diagonalizable, but it may not be the case of $\mathbb{A}$. Anyway, if the two sates $\boldsymbol{u}_L, \boldsymbol{u}_R$ are close enough so that $\mathbb{A}$ is diagonalizable, and if the Riemann problem with left and right states $(\boldsymbol{u}_L, \boldsymbol{u}_R)$ has a solution consisting of a single discontinuity (shock or contact for the Euler equations), then

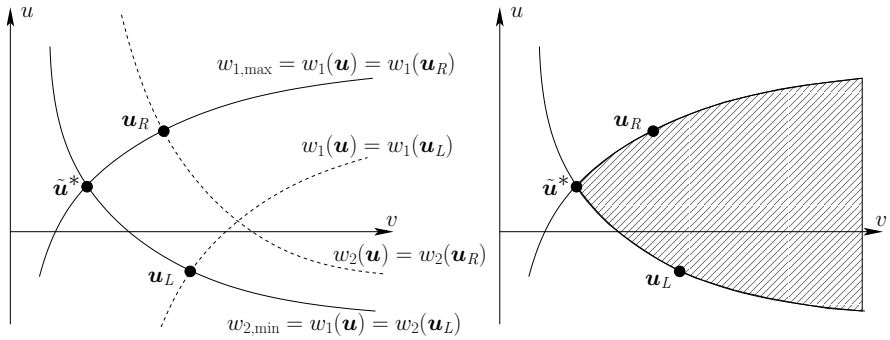

Figure 81.1: Phase space for the p-system with two states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$. Left: definition of $\widetilde{\boldsymbol{u}}^*$ such that $w_1^{\max} = w_1(\widetilde{\boldsymbol{u}}^*)$ and $w_2^{\min} = w_2(\widetilde{\boldsymbol{u}}^*)$. Right: invariant set for the solution to the Riemann problem.

the wave speed of the discontinuity is one of the eigenvalues of $\mathbb{A}$; see, e.g., Bressan [50, §5.2]. In this case, the spectral radius of $\mathbb{A}$ is a guaranteed upper bound of the maximum wave speed. This observation is at the origin of the popularity of the average matrix (sometimes called Roe's matrix). But, the above argument relies on two if's and in general there is no guarantee that the spectral radius of $\mathbb{A}$ is an upper bound of the maximum wave speed in the Riemann problem, as demonstrated in Exercise 81.4. Although it is a common practice in the engineering literature, it is not recommended to use the spectral radius of $\mathbb{A}$ as an ansatz for $\lambda_{\max}(\boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. □

Exercises

Exercise 81.1 (1D approximation). Consider the one-dimensional problem $\partial_t u + \nabla\cdot\boldsymbol{f}(u) = 0$ with $D := (-1, 1)$ and $\boldsymbol{f}(v) := f(v)\boldsymbol{e}_x$. Let $I \in \mathbb{N}$, $I \geq 3$, and consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$ for all $i \in \{1{:}I{-}1\}$, such that $-1 =: x_1 < \cdots < x_I := 1$, with $h_i := x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$. (i) Compute $\boldsymbol{c}_{i,i-1}$ and $\boldsymbol{n}_{i,i-1}$ for all $i \in \{2{:}I\}$, $\boldsymbol{c}_{i,i}$ and m_i for all $i \in \{2{:}I{-}1\}$, and $\boldsymbol{c}_{i,i+1}$ and $\boldsymbol{n}_{i,i+1}$ for all $i \in \{1{:}I{-}1\}$. (ii) Assuming that f is convex, compute $\lambda_{\max}(\boldsymbol{n}_{i,i-1}, \mathsf{U}_i^n, \mathsf{U}_{i-1}^n)$, $\lambda_{\max}(\boldsymbol{n}_{i-1,i}, \mathsf{U}_{i-1}^n, \mathsf{U}_i^n)$, $\lambda_{\max}(\boldsymbol{n}_{i,i+1}, \mathsf{U}_i^n, \mathsf{U}_{i+1}^n)$, and $\lambda_{\max}(\boldsymbol{n}_{i+1,i}, \mathsf{U}_{i+1}^n, \mathsf{U}_i^n)$. (iii) Compute $d_{i,i-1}^n$ and $d_{i,i+1}^n$. (iv) Justify (81.11).

Exercise 81.2 (Symmetry). Let $i \in \mathcal{A}_h^\circ$. (i) Show that $\boldsymbol{c}_{ij} = -\boldsymbol{c}_{ji}$ for all $j \in \mathcal{I}(i)$. (ii) Show that $\lambda_{\max}(\boldsymbol{n}_{ij}, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}\|_{\ell^2} = \lambda_{\max}(\boldsymbol{n}_{ji}, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}\|_{\ell^2}$.

Exercise 81.3 (Average matrix). Let $\mathcal{A} \subset \mathbb{R}^m$ and $\mathbb{f} \in \mathrm{Lip}(\mathcal{A}; \mathbb{R}^{m\times d})$ with components $(\mathbb{f}_{kl})_{k\in\{1:m\},l\in\{1:d\}}$. Let $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathbb{R}^m$ and consider the matrix $\mathbb{A}_{kk'} := \int_0^1 \partial_{v_{k'}}(\mathbb{f}\cdot\boldsymbol{n})_k(\boldsymbol{u}_R + \theta(\boldsymbol{u}_L - \boldsymbol{u}_R))\,\mathrm{d}\theta$. (i) Show that $(\mathbb{f}(\boldsymbol{u}_L) - \mathbb{f}(\boldsymbol{u}_R))\cdot\boldsymbol{n} = \mathbb{A}(\boldsymbol{u}_L - \boldsymbol{u}_R)$. (ii) Assume from now on that $m := 1$ and set $A := \mathbb{A}$, i.e., we are working with scalar equations. Compute A if $u_L \neq u_R$, $\lim_{u_R\to u_L} A$ and $\lim_{u_L\to u_R} A$ assuming that $\mathbb{f}$ is C^1. (iii) Under which conditions do we have $|A| = \lambda_{\max}(\boldsymbol{n}, u_L, u_R)$ if $\mathbb{f}$ is either convex or concave? (*Hint*: see §79.2.) (iv) Take $d_{ij}^n := |A|$ in (81.9) with $\boldsymbol{n} := \boldsymbol{n}_{ij}$, $u_L := \mathsf{U}_i^n$, and $u_R := \mathsf{U}_j^n$. Prove that Theorem 81.8 still holds true if τ is small enough.

Exercise 81.4 (Entropy glitch). Consider the one-dimensional problem $\partial_t u + \nabla\cdot(f(u)\boldsymbol{e}_x) = 0$ with $D := (-1, 1)$ and data $u_0(x) := -1$ if $x \leq 0$ and $u_0(x) := 1$ otherwise. Let $I \in \mathbb{N}\backslash\{0\}$

be an even number, and consider the mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, $i \in \{1{:}I{-}1\}$, such that $-1 =: x_1 < \cdots < x_I := 1$ and $x_{\frac{I}{2}} \le 0 < x_{\frac{I}{2}+1}$. Let $h_i := x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$. (i) Take $d_{ij}^n := \|\boldsymbol{c}_{ij}\|_{\ell^2} |(f(\mathsf{U}_i^n) - f(\mathsf{U}_j^n))/(\mathsf{U}_i^n - \mathsf{U}_j^n)|$ if $\mathsf{U}_i^n \neq \mathsf{U}_j^n$ and $d_{ij}^n := \|\boldsymbol{c}_{ij}\|_{\ell^2} |f'(\mathsf{U}_i^n)|$ otherwise. Prove that Theorem 81.8 still holds true if τ is small enough. (ii) Consider Burgers' flux $\boldsymbol{f}(u) := \frac{1}{2}u^2 \boldsymbol{e}_x$. Take $u_h^0(x) := \sum_{i \in \mathcal{A}_h} \mathsf{U}_i^0 \varphi_i(x)$ with $\mathsf{U}_i^0 := -1$ if $i \le \frac{1}{2}I$ and $\mathsf{U}_i^0 := 1$ if $i \ge \frac{1}{2}I + 1$. Using the above definition of d_{ij}^n, show that the scheme (81.9) gives $u_h^n = u_h^0$ for any $n \ge 0$. Comment on this result.

Chapter 82

Higher-order approximation

The objective of this chapter is to describe techniques that preserve the invariant domain property of the algorithm introduced in Chapter 81 and increase its accuracy in time and space. The argumentation for the time approximation is done for general hyperbolic systems, but the argumentation for the space approximation is done for scalar conservation equations only. The general situation is treated in Chapter 83.

82.1 Higher order in time

Keeping the invariant domain property while increasing the time accuracy can be done by using time discretization methods called *contractive* or *strong stability preserving (SSP)* in the literature. We are mostly going to use the SSP terminology in this chapter. This section is meant to give a brief overview of SSP methods combined with explicit Runge–Kutta (ERK) methods, i.e., SSPRK methods. We refer the reader to Kraaijevanger [205], Ferracina and Spijker [126], Higueras [181], Gottlieb et al. [141] for more detailed reviews.

82.1.1 Key ideas

The key to achieve higher-order accuracy in time is to make convex combinations of forward Euler steps that all have the invariant domain property. More precisely, each time step of a contractive or SSP method is decomposed into substeps that are all forward Euler steps, and the final update is constructed as a convex combination of the intermediate solutions.

Let us motivate the use of SSP methods in the context of the approximation of hyperbolic systems by the algorithm described in (81.18). We introduce the nonlinear operator $L : \mathbb{R}^{m\times I} \to \mathbb{R}^{m\times I}$ s.t. for all $i \in \mathcal{A}_h := \{1{:}I\}$, the component $L(\mathbf{U})_i \in \mathbb{R}^m$ is defined by

$$L(\mathbf{U})_i := \frac{1}{m_i} \sum_{j\in\mathcal{I}(i)} \Big(\mathbb{f}(\mathbf{U}_j)\cdot \boldsymbol{c}_{ij} - d_{ij}(\mathbf{U}_j - \mathbf{U}_i) \Big). \tag{82.1}$$

Recall that the dependence of the graph viscosity d_{ij} on $\mathbf{U}_i, \mathbf{U}_j$ is nonlinear (see (81.19)). Then one step of the algorithm (81.18) consists of setting

$$\mathbf{U}^{n+1} := \mathbf{U}^n + \tau L(\mathbf{U}^n). \tag{82.2}$$

Let $B \subset \mathcal{A}$ be an invariant set. Theorem 81.14 (see also Theorem 81.8 for scalar equations) states that under the CFL condition

$$2\tau \max_{i \in \mathcal{A}_h^\circ} \frac{|d_{ii}^n|}{m_i} < 1, \tag{82.3}$$

where $d_{ii}^n := -\sum_{j \in \mathcal{I}(i) \setminus \{i\}} d_{ij}^n$ and d_{ij}^n is evaluated from $\mathbf{U}_i^n$, $\mathbf{U}_j^n$ as in (81.19), we have

$$[\,\{\mathbf{U}_i^n\}_{i \in \mathcal{A}_h} \subset \mathrm{int}(B)\,] \implies [\,\{\mathbf{U}_i^{n+1}\}_{i \in \mathcal{A}_h} \subset \mathrm{int}(B)\,]. \tag{82.4}$$

That is, $\mathrm{int}(B)$ is invariant under the action of the solution operator $\mathbb{I} + \tau L$. The same property holds true for $\overline{B}$ if $2\tau \max_{i \in \mathcal{A}_h^\circ} \frac{|d_{ii}^n|}{m_i} \le 1$. We want to construct higher-order time-stepping techniques that preserve this property.

To allow for a bit more generality, we consider a finite-dimensional vector space E, a subset $A \subset E$, a time horizon $T > 0$, and a (time-dependent, nonlinear) operator $L : [0,T] \times A \to E$. We are interested in approximating the time-evolution problem $\partial_t u - L(t,u) = 0$. We assume that this problem makes sense (for instance, L continuous in t and Lipschitz in u). We further assume that there exists a convex subset $B \subset A$ and $\tau_* > 0$ such that for all $t \in [0,T]$ and all $s \in [0,\tau_*]$, we have

$$[\,v \in B\,] \implies [\,v + sL(t,v) \in B\,]. \tag{82.5}$$

The time-stepping methods we have in mind to solve the problem $\partial_t u = L(t,u)$ are s-stage ERK schemes where every substep is a convex combination of forward Euler steps. Although the theory of SSPRK methods can be done using the Butcher representation introduced in §69.2.4 for IRK methods and in §78.1 for ERK methods (see, e.g., [205]), for implementation and pedagogical purposes we are going to use a representation introduced in Shu and Osher [259, p. 445]. In this representation (often called (α-β) representation in the SSP literature), every s-stage SSPRK method is defined by two sets of real coefficients α_{ik} and β_{ik} where $i \in \{1\!:\!s\}$ and $k \in \{0\!:\!i\!-\!1\}$ (that is, $1 \le k\!+\!1 \le i \le s$). One also uses the Butcher coefficients $\{c_i\}_{i \in \{1:s\}}$, but to be coherent with the notation used in the SSP literature we shift the indices and define $\gamma_k = c_{k+1}$ for all $k \in \{0\!:\!s\!-\!1\}$. The method proceeds as follows for all $n \in \mathcal{N}_\tau$: Given $u^n \in A$, we first set $w^{(0)} := u^n$, and then we compute $\{w^{(i)}\}_{i \in \{1:s\}}$ by setting

$$w^{(i)} := \sum_{k \in \{0:i-1\}} \alpha_{ik} w^{(k)} + \beta_{ik} \tau L(t_n + \gamma_k \tau, w^{(k)}), \quad \forall i \in \{1\!:\!s\}. \tag{82.6}$$

The update at t_{n+1} is given by $u^{n+1} := w^{(s)}$. The coefficients α_{ik} must satisfy $\sum_{k \in \{0:i-1\}} \alpha_{ik} = 1$ for all $i \in \{1\!:\!s\}$. This is a simple consistency property ensuring that $u^n = w^{(0)} = \ldots = w^{(s)} = u^{n+1}$ whenever $L = 0$. More importantly, for the method to be SSP, the coefficients α_{ik} and β_{ik} must be such that

$$\alpha_{ik} \ge 0, \qquad \beta_{ik} \ge 0, \qquad \text{and} \qquad [\,\alpha_{ik} = 0\,] \implies [\,\beta_{ik} = 0\,], \tag{82.7}$$

for all $1 \le k+1 \le i \le s$. Owing to the implication in (82.7), the computation of $w^{(i)}$ can be rewritten as follows:

$$w^{(i)} := \sum_{k \in \mathcal{K}_i} \alpha_{ik} \Big(w^{(k)} + \alpha_{ik}^{-1} \beta_{ik} \tau L(t_n + \gamma_k \tau, w^{(k)}) \Big), \tag{82.8}$$

where $\mathcal{K}_i := \{k \in \{0\!:\!i\!-\!1\} \mid \alpha_{ik} \ne 0\}$. Since $\sum_{k \in \mathcal{K}_i} \alpha_{ik} = \sum_{k \in \{0:i-1\}} \alpha_{ik} = 1$ and $\alpha_{ik} \ge 0$ by assumption, (82.8) shows that all the intermediate states $w^{(i)}$ are convex combinations of quantities resulting from forward Euler steps. We henceforth set

$$c_{\mathrm{os}} := \inf_{i \in \{1:s\}} \inf_{k \in \mathcal{K}_i} \alpha_{ik} \beta_{ik}^{-1}. \tag{82.9}$$

Notice that it may happen that $\beta_{ik} = 0$ and $\alpha_{ik} \neq 0$. In this case, one sets conventionally $\alpha_{ik}\beta_{ik}^{-1} = \infty$. The following theorem is the main result of this section.

Theorem 82.1 (Shu–Osher). *Let the SSPRK method be defined in* (82.6) *with coefficients satisfying* (82.7). *Let $B \subset E$ be a convex set and assume that there is τ_* such that* (82.5) *holds true. Let c_{os} be defined in* (82.9). *Then the following holds true for all $\tau \le c_{\mathrm{os}}\tau_*$:*

$$[\, u^n \in B \,] \implies [\, u^{n+1} \in B \,]. \tag{82.10}$$

Proof. This result has been established in a slightly different form in Shu and Osher [259, Prop. 2.1] without invoking convexity explicitly. Assume that $u^n \in B$. Let us prove by induction that $w^{(i)} \in B$ for all $i \in \{0{:}s\}$. The assertion holds true for $i = 0$ since $w^{(0)} := u^n$. Consider now $i \in \{1{:}s\}$ and assume that $w^{(k)} \in B$ for all $k \in \{0{:}i{-}1\}$. Setting $z^{(i,k)} := w^{(k)} + \alpha_{ik}^{-1}\beta_{ik}\tau L(t_n + \gamma_k\tau, w^{(k)})$ for all $k \in \mathcal{K}_i$, (82.8) implies that $w^{(i)} = \sum_{k\in\mathcal{K}_i} \alpha_{ik} z^{(i,k)}$. The assumption (82.5) and $\tau \le c_{\mathrm{os}}\tau_*$, together with the definition (82.9) of c_{os}, are sufficient to ascertain that $z^{(i,k)} \in B$ for all $k \in \mathcal{K}_i$. Since $\sum_{k\in\mathcal{K}_i} \alpha_{ik} = \sum_{k\in\{0:i-1\}} \alpha_{ik} = 1$ and $\alpha_{ik} \ge 0$ by assumption, the convexity of the set B implies that $w^{(i)} \in B$. Hence, $w^{(i)} \in B$ for all $i \in \{0{:}s\}$. The statement for $i = s$ is the assertion. □

Example 82.2 (Application). If it can be asserted that there exists τ_*, uniform w.r.t. n, s.t. (82.3) holds true for all $\tau \le \tau_*$, then Theorem 82.1 can be applied for all $n \ge 0$. For example, the reader is invited to verify that uniformity w.r.t. n can be proved for nonlinear scalar equations, the p-system, and the shallow-water equations. It is an open (very hard) question for the compressible Euler equations that is directly related to determining whether the velocity $\boldsymbol{v} := \rho^{-1}\boldsymbol{m}$ stays bounded in time. If the independence of τ_* w.r.t. n is unknown, one can still apply SSPRK methods, but in this case the conclusion of Theorem 82.1 holds true provided that the time step is small enough for all the forward Euler updates in (82.6) to remain in $\mathrm{int}(B)$. For instance, denoting by $\{d_{jj'}^{(ik),n}\}_{j,j'\in\mathcal{A}_h}$ the graph viscosities associated with the forward Euler step $\mathbf{W}^{(k)} + \alpha_{ik}^{-1}\beta_{ik}\tau L(\mathbf{W}^{(k)})$ in (82.6), one must make sure that $2\alpha_{ik}^{-1}\beta_{ik}\tau \max_{j\in\mathcal{A}_h^\circ} \frac{|d_{jj}^{(ik),n}|}{m_j} < 1$ for all $1 \le k+1 \le i \le s$. □

Remark 82.3 (Nonnegativity of β_{ik}). The assumption $\beta_{ik} \ge 0$ can be lifted by using a trick first suggested in Shu [258, Rem. 3.2]. If $\beta_{ik} < 0$, one additionally assumes that one can construct a consistent perturbation of L, say $\widetilde{L} : [0,T]\times A \to E$, such that $s \in B$ implies that $v + s\widetilde{L}(t,v) \in B$ for all $t \in [0,T]$ and all $s \in [-\tau_*, 0]$. Then the computation of $w^{(i)}$ in (82.6) becomes $w^{(i)} = \sum_{\{\beta_{ik}>0\}} \alpha_{ik} w^{(k)} + \beta_{ik}\tau L(t_n + \gamma_k\tau, w^{(k)}) + \sum_{\{\beta_{ik}<0\}} \alpha_{ik} w^{(k)} + \beta_{ik}\tau\widetilde{L}(t_n + \gamma_k\tau, w^{(k)})$. The reader is invited to verify that Theorem 82.1 still holds true with this modification and c_{os} redefined as $c_{\mathrm{os}} := \inf_{i\in\{1:s\}} \inf_{k\in\mathcal{K}_i} \alpha_{ik}|\beta_{ik}|^{-1}$. We refer the reader to Gottlieb et al. [140], Ruuth and Spiteri [245] for further details. For instance, with the operator L defined in (82.1), the operator $\widetilde{L}$ is s.t.

$$\widetilde{L}(\mathbf{U})_i = \frac{1}{m_i} \sum_{j\in\mathcal{I}(i)} \Big(\mathbb{f}(\mathbf{U}_j)\cdot\boldsymbol{c}_{ij} + d_{ij}(\mathbf{U}_j - \mathbf{U}_i) \Big),$$

i.e., one changes the sign of the graph viscosity contribution. □

Remark 82.4 (Computational efficiency). Let us consider two SSPRK methods consisting of s_1 and s_2 stages and with coefficients $c_{\mathrm{os},1}$ and $c_{\mathrm{os},2}$, respectively. Assume that both methods have the same order of accuracy. Considering that the amount of work to compute u^{n+1} from u^n is proportional to the number of stages, one could be led to conclude that the method with the smallest number of stages is the most efficient. This is always the case if $c_{\mathrm{os},1} = c_{\mathrm{os},2}$, but the situation is different if $c_{\mathrm{os},1} \neq c_{\mathrm{os},2}$. Let T be the final time one wants to reach, and let

$N_1 := \lceil T/c_{\mathrm{os},1}\tau_* \rceil$ and $N_2 := \lceil T/c_{\mathrm{os},2}\tau_* \rceil$, i.e., N_1, N_2 are the total numbers of time steps that are necessary to reach T for each method. Note that $N_1/N_2 \approx c_{\mathrm{os},2}/c_{\mathrm{os},1}$. Since the amount of work required by method $l \in \{1,2\}$ is proportional to $s_l N_l$, the ratio of work for method 1 to that for method 2 is $s_1 N_1/(s_2 N_2) \approx s_1 c_{\mathrm{os},2}/(s_2 c_{\mathrm{os},1})$. This leads us to define the *efficiency coefficient* $c_{\mathrm{eff}} := \frac{c_{\mathrm{os}}}{s}$, and we conclude that the larger this coefficient, the more efficient the method. We refer the reader to Gottlieb et al. [141] for a literature review of these questions. □

82.1.2 Examples

SSPRK methods composed of s stages and accurate to order p are often denoted by SSPRK(s,p). We now go through some examples of such methods.

Example 82.5 (SSPRK(2,2)). Heun's method, which is a second-order accurate, two-stage ERK method, is SSP. With obvious notation, it has the following (α-β) tableau (we also include the values of the coefficients $\{\gamma_k\}_{k\in\{0:s-1\}}$) and it can be implemented as follows:

α		β		γ
1		1		0
$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	1

$$w^{(1)} := u^n + \tau L(t_n, u^n),$$
$$w^{(2)} := \tfrac{1}{2}u^n + \tfrac{1}{2}\big(w^{(1)} + \tau L(t_n + \tau, w^{(1)})\big),$$

leading to $c_{\mathrm{os}} = 1$. The midpoint rule (see (82.15) below), which is another second-order accurate two-stage ERK method, is not SSP. □

Example 82.6 (SSPRK(3,3), SSPRK(4,3)). The following third-order accurate three-stage ERK method is SSP:

α			β			γ
1			1			0
$\frac{3}{4}$	$\frac{1}{4}$		0	$\frac{1}{4}$		1
$\frac{1}{3}$	0	$\frac{2}{3}$	0	0	$\frac{2}{3}$	$\frac{1}{2}$

$$w^{(1)} := u^n + \tau L(t_n, u^n),$$
$$w^{(2)} := \tfrac{3}{4}u^n + \tfrac{1}{4}\big(w^{(1)} + \tau L\big(t_n + \tau, w^{(1)}\big)\big),$$
$$w^{(3)} := \tfrac{1}{3}u^n + \tfrac{2}{3}\big(w^{(2)} + \tau L\big(t_n + \tfrac{1}{2}\tau, w^{(2)}\big)\big),$$

leading to $c_{\mathrm{os}} = 1$. The following third-order accurate four-stage ERK method is also SSP:

α				β				γ
1				$\frac{1}{2}$				0
0	1			0	$\frac{1}{2}$			$\frac{1}{2}$
$\frac{2}{3}$	0	$\frac{1}{3}$		0	0	$\frac{1}{6}$		1
0	0	0	1	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$

$$w^{(1)} := u^n + \tau L(t_n, u^n),$$
$$w^{(2)} := w^{(1)} + \tfrac{1}{2}\tau L(t_n + \tfrac{1}{2}\tau, w^{(1)}),$$
$$w^{(3)} := \tfrac{2}{3}u^n + \tfrac{1}{3}\big(w^{(2)} + \tfrac{1}{2}\tau L(t_n + \tau, w^{(2)})\big),$$
$$w^{(4)} := w^{(3)} + \tfrac{1}{2}\tau L(t_n + \tfrac{1}{2}\tau, w^{(3)}),$$

leading to $c_{\mathrm{os}} = 2$. The efficiency coefficients of SSPRK$(3,3)$ and SSPRK$(4,3)$ are $\frac{1}{3}$ and $\frac{1}{2}$, respectively (see Remark 82.4). This suggests that SSPRK$(4,3)$ is actually more computationally efficient than SSPRK$(3,3)$. □

Example 82.7 (SSPRK(5,4)). The following (α-β) tableau describes a fourth-order accurate,

five-stage SSPRK method (see Ruuth [244, Tab. 3]):

α				
1				
0.444370493651235	0.555629506348765			
0.620101851488403	0	0.379898148511597		
0.178079954393132	0	0	0.821920045606868	
0	0	0.517231671970585	0.096059710526147	0.386708617503269
β				
0.391752226571890				
0	0.368410593050371			
0	0	0.251891774271694		
0	0	0	0.544974750228521	
0	0	0	0.063692468666290	0.226007483236906
γ				
0	0.391752226571889	0.586079689311541	0.474542363121399	0.935010630967651

Here, we have $c_{\mathrm{os}} \approx 1.508$. Let us also mention that, as shown in Kraaijevanger [205, Thm. 9.6], there is no SSPRK(4,4) method with $\beta_{ik} \geq 0$ for all $1 \leq k+1 \leq i \leq s$. □

Remark 82.8 (Optimality). Following Theorem 82.1, an optimal SSPRK method is one that maximizes the coefficient c_{os} defined in (82.9). Given a pair (s,p), a natural question is to find an optimal SSPRK(s,p) method. An answer to this question has been given by Kraaijevanger [205], Ferracina and Spijker [126], Higueras [181], Ruuth [244] using fundamental tools on RK methods developed in [205]. Referring to Remark 82.9 for more details, we just comment here on the optimality of the above SSPRK methods. The entire family of optimal SSPRK(s,2) methods is described in [205, Thm. 9.3]. The optimality of the SSPRK(3,3) method is shown in [205, Thm. 9.4], that of the SSPRK(4,3) method in [205, Thm. 9.5], and that of the SSPRK(5,4) method in [205, p. 522] (rediscovered in Spiteri and Ruuth [265]). □

82.1.3 Butcher tableau versus (α-β) representation

Recall from §78.1 that explicit Runge–Kutta (ERK) methods are usually identified by their *Butcher tableau* composed of a (strictly lower-triangular) matrix $\mathcal{A} = (a_{ij})_{i,j\in\{1:s\}} \in \mathbb{R}^{s\times s}$ and a vector $b := (b_i)_{i\in\{1:s\}} \in \mathbb{R}^s$, where s is the number of stages. The conventional representation of ERK methods is as follows:

$$\begin{array}{c|ccccc} c_1 & 0 & & & & \\ c_2 & a_{21} & 0 & & & \\ c_3 & a_{31} & a_{32} & 0 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \tag{82.11}$$

In all the methods considered in this chapter, we take $c_1 := 0$. For all $n \in \mathcal{N}_\tau$, given $u^n \in A$, the update $u^{n+1} \in A$ is obtained by first setting $u^{n,1} := u^n$, then by computing the sequence $\{u^{n,i}\}_{i\in\{1:s\}}$ s.t.

$$u^{n,i} := u^n + \tau \sum_{j\in\{1:i-1\}} a_{ij} L(t_n + c_j\tau, u^{n,j}), \qquad \forall i \in \{2:s\}, \tag{82.12}$$

and finally by setting $u^{n+1} := u^n + \sum_{i\in\{1:s\}} \tau b_i L(t_n + c_i\tau, u^{n,i})$.

The connection between the representation (82.12) given by the Butcher tableau and the representation (82.6) given by the (α-β) tableau has been investigated thoroughly in Ferracina and

Spijker [126, Thm. 2.2], Higueras [181, Prop. 2.1&2.8]. It is shown therein that given an (α-β) representation, there is a unique associated Butcher representation (see Exercise 82.1). For instance, the Butcher tableaux of the SSPRK(2,2), SSPRK(3,3), and SSPRK(4,3) methods introduced above are

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{2} & \frac{1}{2} & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2} \end{array} \tag{82.13}$$

Recall that $\gamma_k := c_{k+1}$ for all $k \in \{0{:}s{-}1\}$. The Butcher tableau of the SSPRK$(5,4)$ method is as follows:

$b_1 = 0.14681187608478644956$	$a_{21} = c_2$
$b_2 = 0.24848290944497614757$	$a_{31} = 0.21766909626116921036$
$b_3 = 0.10425883033198029567$	$a_{32} = 0.36841059305037202075$
$b_4 = 0.27443890090134945681$	$a_{41} = 0.08269208665781075441$
$b_5 = 0.22600748323690765039$	$a_{42} = 0.13995850219189573938$
$c_2 = 0.39175222657188905833$	$a_{43} = 0.25189177427169263984$
$c_3 = 0.58607968931154123111$	$a_{51} = 0.06796628363711496324$
$c_4 = 0.47454236312139913362$	$a_{52} = 0.11503469850463199467$
$c_5 = 0.93501063096765159845$	$a_{53} = 0.20703489859738471851$
	$a_{54} = 0.54497475022851992204$

(82.14)

Conversely, given a Butcher representation, one can construct infinitely many (α-β) representations. If all of them fail to deliver coefficients α_{ik} and β_{ik} satisfying (82.7), the RK method is not SSP. For instance, the midpoint rule, which is a second-order two-stage ERK method defined by the Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array} \qquad \begin{aligned} u^{n,1} &:= u^n \\ u^{n,2} &:= u^n + \tfrac{1}{2}\tau L(t_n, u^{n,1}), \\ u^{n+1} &:= u^n + \tau L(t_n + \tfrac{1}{2}\tau, u^{n,2}), \end{aligned} \tag{82.15}$$

is not SSP. Indeed, one must have $w^{(1)} = u^{n,2}$, which gives $\alpha_{10} = 1$ and $\beta_{10} = \frac{1}{2}$. One must also have $w^{(2)} = u^{n+1} = u^n + \tau L(t_n + \frac{1}{2}\tau, u^{n,2})$, which implies that $\alpha_{20} + \alpha_{21} = 1$, $\beta_{10}\alpha_{21} + \beta_{20} = 0$, $\beta_{21} = 1$, and the second equality requires that either $\beta_{20} < 0$ or $\alpha_{21} < 0$.

Remark 82.9 (Absolute monotonicity of RK methods). Following [205, Def. 2.4], an s-stage RK method with coefficients $(\mathcal{A}, b)$ is said to be absolutely monotone at a given point $\xi \le 0$ if $I - \xi\mathcal{A}$ is nonsingular, $1 + \xi b^\mathsf{T}(I - \xi\mathcal{A})^{-1}e \ge 0$, $\mathcal{A}(I - \xi\mathcal{A})^{-1} \ge 0$, $b^\mathsf{T}(I - \xi\mathcal{A})^{-1} \ge 0$, and $(I - \xi\mathcal{A})^{-1}e \ge 0$, where $e := (1, \dots, 1)^\mathsf{T} \in \mathbb{R}^s$, and the vector inequalities are understood componentwise. Furthermore, the method is said to be absolutely monotone on a given set $S \subset \mathbb{R}$ if it is absolutely monotone at each $\xi \in S$. The radius of absolute monotonicity $R(\mathcal{A}, b)$ is defined by

$$R(\mathcal{A}, b) := \sup\{r \mid r \ge 0 \text{ and } (\mathcal{A}, b) \text{ is absolutely monotone on } [-r, 0]\}.$$

We set $R(\mathcal{A}, b) := 0$ if there is no $r > 0$ such that $(\mathcal{A}, b)$ is absolutely monotone on $[-r, 0]$. Under appropriate assumptions on the operator L, it is shown in [205, Thm. 5.4], Ferracina and Spijker [126, Thm. 3.4], Higueras [181, Prop. 2.7] that the RK method is SSP if $R(\mathcal{A}, b) > 0$ and in this case, the largest possible coefficient c_{os} is $R(\mathcal{A}, b)$. □

82.2 Higher order in space for scalar equations

We revisit in this section the method introduced in §81.1 and make it higher-order accurate in space (it is at least informally second-order accurate). The techniques presented in this section only apply to scalar conservation equations. The general case is addressed in Chapter 83. The material presented here is adapted from Guermond and Popov [158].

82.2.1 Heuristic motivation and preliminary result

The idea we have in mind is to reduce the graph viscosity in regions where the solution is far from a local extremum and keep it first-order in regions where the maximum principle is in danger of being violated. To formalize this idea, we change the notation and denote by $d_{ij}^{\mathrm{L},n}$ the graph viscosity defined in (81.10). We have added the superscript "L" to mean "low-order". We introduce for all $n \geq 0$ a collection of weights $\psi_i^n \in [0,1]$ for all $i \in \mathcal{A}_h$, and we define the high-order graph viscosity as follows:

$$d_{ij}^n := d_{ij}^{\mathrm{L},n} \max(\psi_i^n, \psi_j^n), \qquad \forall j \in \mathcal{I}(i)\backslash\{i\}, \tag{82.16}$$

with the usual convention that $d_{ii}^n := -\sum_{j\in\mathcal{I}(i)\backslash\{i\}} d_{ij}^n$. The high-order approximate solution $\mathbf{U}_i^{n+1}$ for all $i \in \mathcal{A}_h$ is still defined in (81.9), that is,

$$m_i \frac{\mathsf{U}_i^{n+1} - \mathsf{U}_i^n}{\tau} + \sum_{j\in\mathcal{I}(i)} \Big(\boldsymbol{f}(\mathsf{U}_j^n)\cdot \boldsymbol{c}_{ij} - d_{ij}^n(\mathsf{U}_j^n - \mathsf{U}_i^n) \Big) = 0, \tag{82.17}$$

but the graph viscosity is now defined in (82.16). Note that the mass matrix is still lumped. The question that we investigate in this section is how to choose the weights ψ_i^n so that U_i^{n+1} satisfies the same local maximum principle as in Theorem 81.8, that is,

$$\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}], \quad \mathsf{U}_i^{\mathrm{m},n} := \min_{j\in\mathcal{I}(i)} \mathsf{U}_j^n, \quad \mathsf{U}_i^{\mathrm{M},n} := \max_{j\in\mathcal{I}(i)} \mathsf{U}_j^n. \tag{82.18}$$

We define the local CFL number based on the low-order viscosity $d_{ij}^{\mathrm{L},n}$ s.t.

$$\gamma_i^n := \frac{2\tau |d_{ii}^{\mathrm{L},n}|}{m_i}, \qquad \forall i \in \mathcal{A}_h^\circ, \tag{82.19}$$

where $d_{ii}^{\mathrm{L},n} := -\sum_{j\in\mathcal{I}(i)\backslash\{i\}} d_{ij}^{\mathrm{L},n}$. If all the weights ψ_i^n are equal to one, Theorem 81.8 implies that (82.18) holds under the CFL condition $\max_{i\in\mathcal{A}_h^\circ} \gamma_i^n \leq 1$.

We now establish a result that will be useful to design weights ψ_i^n that are as small as possible but are large enough so that (82.18) is still satisfied (possibly under a tighter CFL condition than $\max_{i\in\mathcal{A}_h^\circ} \gamma_i^n \leq 1$.) We define the gap parameter $\theta_i^n \in [0,1]$ s.t.

$$\theta_i^n := \frac{\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}} \ \text{ if } \mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0, \qquad \theta_i^n := \frac{1}{2} \ \text{ otherwise.} \tag{82.20}$$

This definition implies that $\mathsf{U}_i^n = \theta_i^n \mathsf{U}_i^{\mathrm{M},n} + (1-\theta_i^n)\mathsf{U}_i^{\mathrm{m},n}$. We also define

$$\gamma_i^{+,n} := \frac{2\tau}{m_i} \sum_{j\in\mathcal{I}(i^+)} d_{ij}^{\mathrm{L},n}, \qquad \gamma_i^{-,n} := \frac{2\tau}{m_i} \sum_{j\in\mathcal{I}(i^-)} d_{ij}^{\mathrm{L},n}, \tag{82.21}$$

with the subsets $\mathcal{I}(i^+) := \{j \in \mathcal{I}(i) \mid \mathsf{U}_i^n < \mathsf{U}_j^n\}$, $\mathcal{I}(i^-) := \{j \in \mathcal{I}(i) \mid \mathsf{U}_j^n < \mathsf{U}_i^n\}$, and $\gamma_i^{\pm,n}$ is conventionally set to zero if $\mathcal{I}(i^\pm)$ is empty.

Lemma 82.10 (Gap estimates). *Let $n \geq 0$ and $i \in \mathcal{A}_h^\circ$. Assume that $\gamma_i^n < 1$ and $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$. Define the real numbers*

$$\delta_i^{\mathrm{M},n} := \big((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{-,n}\big), \tag{82.22a}$$

$$\delta_i^{\mathrm{m},n} := \big(\theta_i^n(1-\gamma_i^n) - (1-\theta_i^n)(1-\psi_i^n)\tfrac{1}{2}\gamma_i^{+,n}\big). \tag{82.22b}$$

Let U_i^{n+1} be given by (82.17)*. Then we have*

$$\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n} + (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\delta_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n} - (\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n)\delta_i^{\mathrm{M},n}]. \tag{82.23}$$

Proof. See Exercise 82.4 and [158, Lem. 4.1]. □

Lemma 82.10 gives an estimate on the gaps between U_i^{n+1} and the two extreme values $\mathsf{U}_i^{\mathrm{m},n}$, $\mathsf{U}_i^{\mathrm{M},n}$. We are going to use this lemma in §82.2.2 and §82.2.3 to devise ways to take the weights ψ_i^n as small as possible while ensuring that $\delta_i^{\mathrm{M},n} \geq 0$ and $\delta_i^{\mathrm{m},n} \geq 0$, possibly under a tighter CFL condition than $\max_{i\in\mathcal{A}_h^\circ}\gamma_i^n \leq 1$. The statement (82.23) will then imply that the local maximum principle (82.18) is satisfied.

82.2.2 Smoothness-based graph viscosity

The technique considered in this section is based on a measure of the local smoothness of the solution. Assuming that $\mathsf{U}_i^{\mathrm{m},n} \neq \mathsf{U}_i^{\mathrm{M},n}$, we introduce the real numbers

$$\alpha_i^n := \frac{\left|\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\beta_{ij}(\mathsf{U}_j^n - \mathsf{U}_i^n)\right|}{\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|} \in [0,1], \tag{82.24}$$

for all $i \in \mathcal{A}_h$ and all $n \geq 0$, where the real numbers β_{ij} are assumed to be nonnegative and not all equal to zero (these numbers should not be confused with the parameters in the SSPRK method). The idea is to define the weights ψ_i^n in (82.16) as

$$\psi_i^n := \psi(\alpha_i^n), \quad \psi \in \mathrm{Lip}([0,1];[0,1]), \quad \psi(1) = 1. \tag{82.25}$$

Whenever U_i^n is a local extremum, the definition (82.24) implies that $\alpha_i^n = 1$, so that $\psi_i^n = \psi(1) = 1$, i.e., the graph viscosity is taken equal to the low-order graph viscosity. This can be a desirable feature since the method degenerates to first order at extrema to avoid violating the local maximum principle. In smooth regions away from local extrema, u_h^n is close to being linear over the support of φ_i, and one would like to take α_i^n as small as possible. One idea is to design the coefficients β_{ij} so that $\alpha_i^n = 0$ when u_h^n is linear. Then, when u_h^n is close to being linear, one expects that the numerator of (82.24) behaves like $h^2\|D^2u(\boldsymbol{\xi},t_n)\|_{\ell^2(\mathbb{R}^{d\times d})}$ at some point $\boldsymbol{\xi}$, whereas the denominator behaves like $h\|\nabla u(\boldsymbol{\zeta},t_n)\|_{\ell^2(\mathbb{R}^d)}$ at some point $\boldsymbol{\zeta}$. In these conditions, $\alpha_i^n \approx h\|D^2u(\boldsymbol{\xi},t_n)\|_{\ell^2(\mathbb{R}^{d\times d})}/\|\nabla u(\boldsymbol{\zeta},t_n)\|_{\ell^2(\mathbb{R}^d)}$, i.e., α_i^n is of order h/ℓ_D (ℓ_D is a characteristic length scale of D, e.g., $\ell_D := \mathrm{diam}(D)$). This makes the method (informally) second-order consistent in smooth regions away from local extrema.

If we take $\psi := 1$, the graph viscosity is equal to the low-order graph viscosity, and we know from Theorem 81.8 that the local maximum principle (82.18) is satisfied under the CFL condition $\max_{i\in\mathcal{A}_h^\circ}\gamma_i^n \leq 1$. In the following result, we quantify how tighter the CFL condition must be to preserve the local maximum principle (82.18) while allowing the weights ψ_i^n to take values in $[0,1]$. Let us denote by $\beta_i^{\mathrm{m}} := \min_{j\in\mathcal{I}(i)}\beta_{ij}$ and $\beta_i^{\mathrm{M}} := \max_{j\in\mathcal{I}(i)}\beta_{ij}$, and suppose that there exists $\beta^\sharp \in (0,\infty)$ such that for all $h \in \mathcal{H}$,

$$0 < \beta_{ij}, \quad \forall i \in \mathcal{A}_h,\ \forall j \in \mathcal{I}(i), \qquad \max_{i\in\mathcal{A}_h}\frac{\beta_i^{\mathrm{M}}}{\beta_i^{\mathrm{m}}} \leq \beta^\sharp. \tag{82.26}$$

Let us set $c_\sharp := \beta^\sharp \max_{i\in\mathcal{A}_h} \mathrm{card}(\mathcal{I}(i))$ (notice that this number is uniformly bounded w.r.t. $h \in \mathcal{H}$).

Theorem 82.11 (Maximum principle with tighter CFL). *Let $\psi \in \mathrm{Lip}([0,1];[0,1])$ be s.t. $\psi(1) = 1$ and let k_ψ be the Lipschitz constant of ψ. Set $\psi_i^n := \psi(\alpha_i^n)$ for all $i \in \mathcal{A}_h$ and all $n \geq 0$, with α_i^n defined in* (82.24) *and the coefficients β_{ij} satisfying the assumptions* (82.26)*. Then the scheme* (82.17) *with d_{ij}^n defined in* (82.16) *satisfies the local maximum principle* (82.18) *under the tighter* CFL condition $\max_{i\in\mathcal{A}_h^\circ} \gamma_i^n \leq \frac{1}{1+k_\psi c_\sharp}$.

Proof. Notice first that if $\mathsf{U}_i^{\mathrm{M},n} = \mathsf{U}_i^{\mathrm{m},n}$, then $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ irrespective of the value of d_{ij}^n, which proves the local maximum principle. Let us assume now that $\mathsf{U}_i^{\mathrm{M},n} \neq \mathsf{U}_i^{\mathrm{m},n}$. If $\theta_i^n = \frac{\mathsf{U}_i^n - \mathsf{U}_i^{\mathrm{m},n}}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}} \in \{0,1\}$, then $\mathsf{U}_i^n \in \{\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}\}$. In this case, (82.24) implies that $\alpha_i^n = 1$ and $\psi_i^n = \psi(\alpha_i^n) = 1$. Using (82.16), we infer that $d_{ij}^n = d_{ij}^{\mathrm{L},n} \max(1, \psi(\alpha_j^n)) = d_{ij}^{\mathrm{L},n}$ for all $j \in \mathcal{I}(i)$. This means that U_i^n coincides with the low-order solution, and since the tighter CFL condition implies the usual CFL condition (81.12), i.e., $\max_{i\in\mathcal{A}_h^\circ} \gamma_i^n \leq 1$, we infer from Theorem 81.8 that $\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. Finally, let us assume that $\theta_i^n \in (0,1)$. Letting $s_i^\pm := \sum_{j\in\mathcal{I}(i^\pm)} \beta_{ij} |\mathsf{U}_j^n - \mathsf{U}_i^n| \geq 0$ and since $-|s_i^+ - s_i^-| \leq s_i^+ - s_i^-$ and $\mathcal{I}(i^+) \subset \mathcal{I}(i)$, we have

$$\begin{aligned}
1 - \alpha_i^n &= 1 - \frac{|s_i^+ - s_i^-|}{s_i^+ + s_i^-} \leq 2\frac{s_i^+}{s_i^+ + s_i^-} \\
&= 2\frac{\sum_{j\in\mathcal{I}(i^+)} \beta_{ij}(\mathsf{U}_j^n - \mathsf{U}_i^n)}{\sum_{j\in\mathcal{I}(i)} \beta_{ij}|\mathsf{U}_j^n - \mathsf{U}_i^n|} \leq 2\frac{\sum_{j\in\mathcal{I}(i^+)} \beta_{ij}(\mathsf{U}_j^{\mathrm{M},n} - \mathsf{U}_i^n)}{\beta_i^m|\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n| + \beta_i^m|\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^n|} \\
&\leq 2\frac{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^n}{\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n}} \frac{\beta_i^{\mathrm{M}}}{\beta_i^{\mathrm{m}}} \mathrm{card}(\mathcal{I}(i^+)) \leq 2c_\sharp(1 - \theta_i^n).
\end{aligned}$$

The last inequality is a consequence of $c_\sharp \geq \frac{\beta_i^{\mathrm{M}}}{\beta_i^{\mathrm{m}}} \mathrm{card}(\mathcal{I}(i))$ for all $i \in \mathcal{A}_h$ with $\beta_i^{\mathrm{m}} := \min_{j\in\mathcal{I}(i)} \beta_{ij}$ and $\beta_i^{\mathrm{M}} := \max_{j\in\mathcal{I}(i)} \beta_{ij}$. Likewise, using that $-|s_i^+ - s_i^-| \leq s_i^- - s_i^+$ and $\mathcal{I}(i^-) \subset \mathcal{I}(i)$, we infer that

$$1 - \alpha_i^n \leq 2c_\sharp\theta_i^n.$$

Let k_ψ be the Lipschitz constant of ψ. Then $1 - \psi(\alpha_i^n) = \psi(1) - \psi(\alpha_i^n) \leq k_\psi(1 - \alpha_i^n) \leq 2k_\psi c_\sharp \min(\theta_i^n, 1 - \theta_i^n)$. Recall the real numbers $\delta_i^{\mathrm{m},n}$ and $\delta_i^{\mathrm{M},n}$ defined in (82.22). Since $\theta_i^n \in (0,1)$ and $\gamma_i^{-,n} \leq \gamma_i^n$, we infer that

$$\begin{aligned}
\delta_i^{\mathrm{M},n} &= (1 - \theta_i^n)(1 - \gamma_i^n) - \theta_i^n(1 - \psi(\alpha_i^n))\tfrac{1}{2}\gamma_i^{-,n} \\
&\geq (1 - \theta_i^n)(1 - \gamma_i^n) - k_\psi c_\sharp \theta_i^n(1 - \theta_i^n)\gamma_i^n \\
&\geq (1 - \theta_i^n)(1 - (1 + k_\psi c_\sharp \theta_i^n)\gamma_i^n) \geq 0,
\end{aligned}$$

provided $\gamma_i^n \leq \frac{1}{1+k_\psi c_\sharp}$. Similarly, provided again that $\gamma_i^n \leq \frac{1}{1+k_\psi c_\sharp}$, we have

$$\begin{aligned}
\delta_i^{\mathrm{m},n} &= \theta_i^n(1 - \gamma_i^n) - (1 - \theta_i^n)(1 - \psi(\alpha_i^n))\tfrac{1}{2}\gamma_i^{+,n} \\
&\geq \theta_i^n(1 - \gamma_i^n) - k_\psi c_\sharp \theta_i^n(1 - \theta_i^n)\gamma_i^n \\
&\geq \theta_i^n(1 - (1 + k_\psi c_\sharp(1 - \theta_i^n))\gamma_i^n) \geq 0.
\end{aligned}$$

The conclusion follows from Lemma 82.10. □

Remark 82.12 (Choosing ψ and warnings). It is essential that $\psi(1) = 1$ to ensure the maximum principle, and it is important that $\psi(\alpha)$ be as close to zero as possible when $\alpha \in [0,1)$

to increase the accuracy. A good candidate is $\psi(\alpha) := \big(\frac{\max(\alpha-\alpha_0,0)}{1-\alpha_0}\big)^p$ with $\alpha_0 \in [0,1)$ and $p \geq 1$. Numerical tests with $p := 2$ and $\alpha_0 := 0$ are reported in [158, §6], and numerical tests with $p := 4$ are reported in [169, §5]. We also refer the reader to [170, §7] for numerical tests on the shallow-water equations. The reader should be aware though that being maximum principle preserving does not guarantee that the method converges to the entropy solution; see [158] for counterexamples. Numerical experiments show that convergence to the entropy solution is achieved if the flux $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ is either strictly convex or concave for all unit vectors $\boldsymbol{n}$, but this may not be the case where $\boldsymbol{f}(v)\cdot\boldsymbol{n}$ has inflection points. This problem is exacerbated as p grows and α_0 gets close to 1 (note that the Lipschitz constant of ψ grows unboundedly in this case). We refer the reader to [158, §6.2] for more details. □

Remark 82.13 (Literature). The idea of reducing the artificial viscosity by measuring the local smoothness of the solution was originally developed in the finite volume literature (see, e.g., Jameson et al. [195, Eq. (12)], Jameson [194]). Theorem 82.11 is established in Guermond and Popov [158]. The quantity $(\alpha_i^n)^p$, $p \geq 2$, is used in Burman [58] to construct a nonlinear viscosity that yields the maximum principle and convergence to the entropy solution for Burgers' equation in one dimension. In Barrenechea et al. [25, Eq. (2.4)-(2.5)], this idea is combined with a graph diffusion operator inspired from Burman and Ern [61] to solve linear scalar advection-diffusion equations. □

Remark 82.14 (Convergence barrier). The property that $\alpha_i^n = 1$ when U_i^n is a local extremum limits the convergence order of the method in the L^∞-norm. Numerical experiments reported in [158, Tab. 6.1] show that the method is second-order accurate in space in the L^1-norm, but it is only first-order accurate in the L^∞-norm. This phenomenon is similar to what is observed in the finite volume literature for total variation diminishing schemes (TVD). As stated in Harten and Osher [178, p. 280], "the perpetual damping of local extrema determines the cumulative global error of the 'high-order TVD schemes' to be $\mathcal{O}(h)$ in the L^∞-norm, $\mathcal{O}(h^{\frac{3}{2}})$ in the L^2-norm, and $\mathcal{O}(h^2)$ in the L^1-norm." □

We now discuss ways to construct the coefficients β_{ij} to make the scheme (82.17) with d_{ij}^n defined in (82.16) linearity preserving in the following sense.

Definition 82.15 (Linearity preserving method). *Methods such that $d_{ij}^n = 0$ for all $j \in \mathcal{I}(i)$ when u_h^n is linear on the support of φ_i are said to be* linearity preserving.

Recalling the discussion below (82.24), the motivation for such a property is to make the scheme (informally) second-order accurate in smooth regions away from local extrema. Let us start with continuous, piecewise linear Lagrange elements on a one-dimensional nonuniform grid with vertices $\{x_i\}_{i\in\{1:I\}}$. Consider two consecutive cells $[x_{i-1}, x_i]$, $[x_i, x_{i+1}]$ with local meshsizes $h_{i-1} := x_i - x_{i-1}$, $h_i := x_{i+1} - x_i$. Up to the boundary vertices, the support of φ_i is $[x_{i-1}, x_{i+1}]$. If u_h^n is linear over $[x_{i-1}, x_{i+1}]$, then $\mathsf{U}_{i+1}^n \frac{x_i - x_{i-1}}{x_{i+1}-x_{i-1}} + \mathsf{U}_{i-1}^n \frac{x_i - x_{i+1}}{x_{i-1}-x_{i+1}} - \mathsf{U}_i^n$ should be equal to zero. This quantity can be rewritten as $(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n)\frac{|x_i - x_{i-1}|}{h_{i-1}+h_i} + (\mathsf{U}_{i-1}^n - \mathsf{U}_i^n)\frac{|x_i - x_{i+1}|}{h_{i-1}+h_i}$. Hence, in one dimension, it is natural to take

$$\beta_{ij} := \frac{|x_i - x_j|}{h_{i-1} + h_i}, \qquad j \in \{i-1, i+1\}. \tag{82.27}$$

Then $\alpha_i^n = 0$ if u_h^n is linear over $[x_{i-1}, x_{i+1}]$, and the method is linearity preserving if one chooses the function ψ so that $\psi(0) = 0$.

The above argument generalizes to higher dimension by making use of *generalized barycentric coordinates*. Let $\{\varphi_i\}_{i\in\mathcal{A}_h}$ be a basis composed of continuous, piecewise linear Lagrange shape

functions associated with the Lagrange nodes $\{\boldsymbol{z}_i\}_{i\in\mathcal{A}_h}$. Let P_i be the polytope with vertices $\{\boldsymbol{z}_j\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ for all $i\in\mathcal{A}_h^\circ$ (in dimension three, the boundary of P_i is assumed to be formed by a triangulation of the faces formed by the vertices $\{\boldsymbol{z}_j\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ if the cells are not tetrahedra). Note that $\mathrm{supp}(\varphi_i)=P_i$. We say that $\{\omega_{ij}(\boldsymbol{x})\}_{j\in\mathcal{I}(i)\backslash\{i\}}$ is a set of generalized barycentric coordinates in P_i if

$$1=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\omega_{ij}(\boldsymbol{x}),\qquad \boldsymbol{x}=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\omega_{ij}(\boldsymbol{x})\boldsymbol{z}_j,\qquad \forall\boldsymbol{x}\in\mathbb{R}^d, \tag{82.28a}$$

$$\omega_{ij}(\boldsymbol{x})\geq 0,\ \forall j\in\mathcal{I}(i)\backslash\{i\},\qquad \forall\boldsymbol{x}\in P_i. \tag{82.28b}$$

The first identity in (82.28) implies that $u_h^n(\boldsymbol{z}_i)=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\omega_{ij}(\boldsymbol{z}_i)u_h^n(\boldsymbol{z}_i)$. Moreover, if u_h^n is linear over P_i, the second identity implies that $u_h^n(\boldsymbol{z}_i)=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\omega_{ij}(\boldsymbol{z}_i)u_h^n(\boldsymbol{z}_j)$. Recalling that for Lagrange finite elements $u_h^n(\boldsymbol{z}_j)=\mathsf{U}_j^n$, this shows that the quantity $\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\omega_{ij}(\boldsymbol{z}_i)(\mathsf{U}_j^n-\mathsf{U}_i^n)$ is zero when u_h^n is linear in P_i. This argument shows that in this case it is natural to take

$$\beta_{ij}:=\omega_{ij}(\boldsymbol{z}_i),\qquad \forall j\in\mathcal{I}(i)\backslash\{i\}, \tag{82.29}$$

and this makes the method linearity preserving if $\psi(0)=0$.

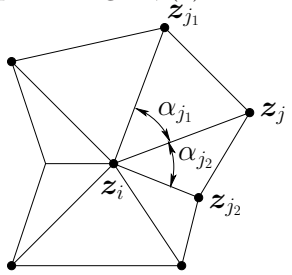

Figure 82.1: Polygon P_i associated with a vertex $\boldsymbol{z}_i$ and notation for the definition of the mean-value coordinates.

There are many ways to construct generalized barycentric coordinates. We refer the reader to Floater [127] for a review of generalized barycentric coordinates on polygons and polyhedra. If P_i is convex, one can use the Wachspress coordinates in dimension two [283] and the technique described in Warren et al. [284] in dimension three. When P_i in not convex, an alternative consists of using the mean-value coordinates proposed in Floater et al. [128]. Suppose that the dimension is two. Let $j\neq i$, $\boldsymbol{z}_{j_1}$, $\boldsymbol{z}_{j_2}$ be the vertices on each side of $\boldsymbol{z}_j$, and α_{j_l} be the unoriented angle $\widehat{\boldsymbol{z}_l\boldsymbol{z}_i\boldsymbol{z}_j}$, $l\in\{1,2\}$ (see Figure 82.1). After setting $\rho_{ij}:=\frac{\tan(\alpha_{j_1}/2)+\tan(\alpha_{j_2}/2)}{\|\boldsymbol{z}_j-\boldsymbol{z}_i\|_{\ell^2}}$, the mean-value coordinates at $\boldsymbol{z}_i$ are defined by

$$\omega_{ij}(\boldsymbol{z}_i):=\frac{\rho_{ij}}{\sum_{k\in\mathcal{I}(i)\backslash\{i\}}\rho_{ik}},\qquad \forall j\in\mathcal{I}(i)\backslash\{i\}. \tag{82.30}$$

If P_i is star-shaped with respect to $\boldsymbol{z}_i$, the angles α_{j_l}, $l\in\{1,2\}$, are less than π, which proves that $\omega_{ij}(\boldsymbol{z}_i)>0$. Note that in our case, P_i is always star-shaped with respect to $\boldsymbol{z}_i$, whether $\widehat{K}$ is the unit simplex or the unit square. A similar construction is available in dimension three; see [127, §8.3].

82.2.3 Greedy graph viscosity

The greedy graph viscosity is another technique to reduce the graph viscosity. It is entirely based on the result of Lemma 82.10, irrespective of any smoothness considerations.

Theorem 82.16 (Greedy graph viscosity). *Let θ_i^n, $\gamma_i^{-,n}$, and $\gamma_i^{+,n}$ be defined in* (82.20)-(82.21) *for all $i \in \mathcal{A}_h$ and all $n \geq 0$. Define the weights ψ_i^n as*

$$\psi_i^n := \max\left(1 - 2(1-\gamma_i^n)\min\left(\frac{1}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}, \frac{1}{\gamma_i^{+,n}}\frac{\theta_i^n}{(1-\theta_i^n)}\right), 0\right), \tag{82.31}$$

with the convention $\psi_i^n := 1$ if $\theta_i^n \in \{0,1\}$. Then the scheme (82.17) *with d_{ij}^n defined in* (82.16) *satisfies the local maximum principle* (82.18) *under the usual CFL condition* $\max_{i\in\mathcal{A}_h^\circ}\gamma_i^n \leq 1$.

Proof. Note first that if $\mathsf{U}_i^{\mathrm{M},n} = \mathsf{U}_i^{\mathrm{m},n}$, then $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ irrespective of the value of d_{ij}^n, which proves the local maximum principle. If $\theta_i^n \in \{0,1\}$, then $\psi_i^n = 1$ so that $d_{ij}^n = d_{ij}^{\mathrm{L},n}\max(1,\psi_j^n) = d_{ij}^{\mathrm{L},n}$ for all $j \in \mathcal{I}(i)\backslash\{i\}$, which again implies that $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$. Finally, let us assume that $\theta_i^n \in (0,1)$. Recall the real numbers $\delta_i^{\mathrm{M},n}$, $\delta_i^{\mathrm{m},n}$ defined in (82.22). The definition of ψ_i^n in (82.31) implies that $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{-,n}}\frac{1-\theta_i^n}{\theta_i^n}$, which in turn gives $\delta_i^{\mathrm{M},n} = (1-\gamma_i^n)(1-\theta_i^n) + \theta_i^n(\psi_i^n - 1)\frac{1}{2}\gamma_i^{-,n} \geq 0$. Similarly, we have $\psi_i^n \geq 1 - 2\frac{1-\gamma_i^n}{\gamma_i^{+,n}}\frac{\theta_i^n}{1-\theta_i^n}$, which gives $\delta_i^{\mathrm{m},n} = (1-\gamma_i^n)\theta_i^n + (\psi_i^n - 1)(1-\theta_i^n)\frac{1}{2}\gamma_i^{+,n} \geq 0$. Lemma 82.10 shows that $\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\mathrm{m},n}, \mathsf{U}_i^{\mathrm{M},n}]$ (see (82.23)). This proves the claim. □

Remark 82.17 (Small CFL number). When the local CFL number γ_i^n is small and U_i^n is not a local extremum, ψ_i^n is close to 0. This shows that the method is greedier as the CFL number decreases, whence the name. □

Remark 82.18 (Min-Max). The greedy graph viscosity based on (82.31) explicitly involves the bounds $\mathsf{U}_i^{\mathrm{m},n}$ and $\mathsf{U}_i^{\mathrm{M},n}$ which are needed to compute θ_i^n (see (82.20)), whereas the smoothness-based graph viscosity with $\psi_i^n := \psi(\alpha_i^n)$ and α_i^n defined in (82.24) does not. □

Exercises

Exercise 82.1 ($(\alpha-\beta)$ vs. Butcher representation). (i) Consider the ERK scheme defined by the Butcher tableau (82.11), i.e., the matrix $\mathcal{A} \in \mathbb{R}^{s\times s}$ and the vector $b \in \mathbb{R}^s$. Consider the matrix $\mathbb{A} := \begin{pmatrix}\mathcal{A} & \mathbf{0}\\ b^\mathsf{T} & 0\end{pmatrix}$ of order $(s+1)$, with $\mathbf{0} := (0,\dots,0)^\mathsf{T} \in \mathbb{R}^s$. Set $u^{(i)} := u^n + \tau\sum_{j\in\{1:i-1\}} a_{ij}k_j$ for all $i \in \{1{:}s\}$, where a_{ij} are the entries of the matrix $\mathcal{A}$. Consider the vectors $\mathsf{U} := (u^{(1)},\dots,u^{(s)},u^{n+1})^\mathsf{T}$ and $\mathsf{F}(\mathsf{U}) := (L(t_n + c_1\tau, u^{(1)}),\dots,L(t_n+c_s\tau,u^{(s)}),0)^\mathsf{T}$. Show that $\mathsf{U} = u^n\mathsf{E} + \tau\mathbb{A}\mathsf{F}(\mathsf{U})$ with $\mathsf{E} := (1,\dots,1)^\mathsf{T} \in \mathbb{R}^{s+1}$. (ii) Consider the scheme defined by the $(\alpha\text{-}\beta)$ representation (82.6) with $\gamma_k := c_{k+1}$ for all $k \in \{0{:}s{-}1\}$. Let $\mathbb{a}$ and $\mathbb{b}$ be the $(s+1)\times(s+1)$ strictly lower triangular matrices with entries $\mathbb{a}_{i+1,k+1} := \alpha_{ik}$, $\mathbb{b}_{i+1,k+1} := \beta_{ik}$ for all $1 \leq k+1 \leq i \leq s$. Show that $(\mathbb{I}-\mathbb{a})\mathsf{E} = \mathsf{E}_1$ with $\mathsf{E}_1 := (1,0,\dots,0)^\mathsf{T} \in \mathbb{R}^{s+1}$. (iii) Consider the vectors $\mathsf{W} := (w^{(0)},\dots,w^{(s)})^\mathsf{T}$, $\mathsf{F}(\mathsf{W}) := (L(t_n+c_1\tau,w^{(0)}),\dots,L(t_n+c_s\tau,w^{(s-1)}),0)^\mathsf{T}$. Show that $\mathsf{W} = u^n\mathsf{E} + \tau(\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}\mathsf{F}(\mathsf{W})$. (iv) Compute the matrices $\mathbb{a}$, $\mathbb{b}$, and $(\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}$ for the SSPRK(2,2) scheme. *Note*: this exercise shows that given the $(\alpha\text{-}\beta)$ representation (82.6), there is only one associated Butcher tableau. But given a Butcher tableau, there may be more than one $(\alpha\text{-}\beta)$ representation since the factorization $\mathbb{A} = (\mathbb{I}-\mathbb{a})^{-1}\mathbb{b}$ may be nonunique.

Exercise 82.2 (Quadratic approximation). (i) Give the expression of the reference shape functions for the Lagrange element $(\widehat{K}, \mathbb{P}_{2,1}, \{\widehat{\sigma}_1,\widehat{\sigma}_2,\widehat{\sigma}_3\})$ where $\widehat{K} := [0,1]$, $\widehat{\sigma}_1(\widehat{p}) := \widehat{p}(0)$, $\widehat{\sigma}_2(\widehat{p}) :=$

$\widehat{p}(\frac{1}{2})$, $\widehat{\sigma}_3(\widehat{p}) := \widehat{p}(1)$. (ii) Compute the reference mass matrix $M_{\widehat{K}}$ with entries $\int_{\widehat{K}} \widehat{\theta}_i(\widehat{x})\widehat{\theta}_j(\widehat{x})\,\mathrm{d}\widehat{x}$. (iii) Compute the lumped reference mass matrix $\overline{M}_{\widehat{K}}$. What should be the sum of the entries of $\overline{M}_{\widehat{K}}$? (iv) Let $D := (0,1)$. Let $N_\mathrm{e} \geq 1$, $I := 2N_\mathrm{e}+1$, and let $0 =: x_1 < \ldots < x_I := 1$. Consider the mesh $\mathcal{T}_h$ composed of the cells $K_m := [x_{2m-1}, x_{2m+1}]$, $\forall m \in \{1{:}N_\mathrm{e}\}$. Let $h_m := x_{2m+1} - x_{2m-1}$. Let $P_2^\mathrm{g}(\mathcal{T}_h)$ be the H^1-conforming space based on $\mathcal{T}_h$ using quadratic polynomials. Give the expression of the global shape functions of $P_2^\mathrm{g}(\mathcal{T}_h)$ associated with the Lagrange nodes $\{x_i\}_{i\in\mathcal{A}_h}$ with $\mathcal{A}_h := \{1{:}I\}$. (v) Give the coefficients of the consistent mass matrix. (vi) Give the coefficients of the lumped mass matrix. What should be the sum of the entries of M^L? (vii) Is it possible to use the above Lagrange basis together with the theory described in §81.1.2 to approximate hyperbolic systems? (viii) Is it possible to apply Corollary 81.9 and Corollary 81.15?

Exercise 82.3 (Quadratic Bernstein approximation). Consider the following reference shape functions on $\widehat{K} := [0,1]$:

$$\widehat{\theta}_1(\widehat{x}) := (1-\widehat{x})^2, \quad \widehat{\theta}_2(\widehat{x}) := 2\widehat{x}(1-\widehat{x}), \quad \widehat{\theta}_3(\widehat{x}) := \widehat{x}^2.$$

(i) Show that $\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3\}$ is a basis of $\mathbb{P}_{2,1}$. Show that these functions satisfy the partition of unity property and that $\widehat{p}(\widehat{x}) = \widehat{p}(0)\widehat{\theta}_1(\widehat{x}) + \widehat{p}(\frac{1}{2})\widehat{\theta}_2(\widehat{x}) + \widehat{p}(1)\widehat{\theta}_3(\widehat{x})$ for all $\widehat{p} \in \mathbb{P}_{1,1}$. (ii)-(viii) Redo Questions (ii)-(viii) of Exercise 82.2 with the above reference shape functions.

Exercise 82.4 (Gap estimates). The objective is to prove Lemma 82.10. (i) Let $\mathsf{U}^{\mathrm{L},n+1}$ be the update given by (81.9) with the low-order graph viscosity d_{ij}^L. Consider the auxiliary states $\overline{\mathsf{U}}_{ij}^n := \frac{1}{2}(\mathsf{U}_j^n + \mathsf{U}_i^n) - (\boldsymbol{f}(\mathsf{U}_j^n) - \boldsymbol{f}(\mathsf{U}_i^n))\cdot\frac{\boldsymbol{c}_{ij}}{2d_{ij}^{\mathrm{L},n}}$ defined in the proof of Theorem 81.8 for all $j \in \mathcal{I}(i)$, and set $\mathsf{U}_i^{*,n} := \frac{1}{\gamma_i^n}\sum_{j\in\mathcal{I}(i)\setminus\{i\}} \frac{2\tau d_{ij}^{\mathrm{L},n}}{m_i}\overline{\mathsf{U}}_{ij}^n$. Show that

$$\mathsf{U}_i^{n+1} = (1-\gamma_i^n)\mathsf{U}_i^n + \gamma_i^n\mathsf{U}_i^{*,n} + \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\setminus\{i\}}(d_{ij}^n - d_{ij}^{\mathrm{L},n})(\mathsf{U}_j^n - \mathsf{U}_i^n).$$

(ii) Using that $\mathsf{U}_{ij}^{*,n} \leq \mathsf{U}_i^{\mathrm{M},n}$, $d_{ij}^n \leq d_{ij}^{\mathrm{L},n}$, and $\mathsf{U}_i^{\mathrm{M},n} - \mathsf{U}_i^{\mathrm{m},n} \neq 0$, show that

$$\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\mathrm{M},n} + (\mathsf{U}_i^{\mathrm{m},n} - \mathsf{U}_i^{\mathrm{M},n})\Big((1-\theta_i^n)(1-\gamma_i^n) - \theta_i^n\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i^-)}(d_{ij}^{\mathrm{L},n} - d_{ij}^n)\Big).$$

(iii) Using that $d_{ij}^n \geq d_{ij}^{\mathrm{L},n}\psi_i^n$ and $\psi_i^n \geq 0$, prove the upper bound in (82.23). (iv) Prove the lower bound in (82.23).

Chapter 83

Higher-order approximation and limiting

This chapter is the continuation of Chapter 82. The objective is to describe techniques for the solution of the hyperbolic system (80.2) that are at least (informally) second-order accurate in space and invariant domain preserving. As seen in Chapter 82, one can make the method more accurate in space by decreasing the first-order graph viscosity. Another technique, which is very efficient when working with nonsmooth data or with solutions with shocks or contact discontinuities, consists of using the consistent mass matrix instead of the lumped mass matrix in the approximation of the time derivative. These two techniques increase the accuracy in space but deliver an update that can step out of the invariant domain. We then show that this defect can be corrected by applying a conservative convex limiting technique. Let us emphasize that the heuristics we have in mind is that limiting should be understood as a light post-processing applied to a method that is already entropy consistent and almost invariant domain preserving. The present material is adapted from Guermond and Popov [158], Guermond et al. [169, 171].

83.1 Higher-order techniques

We present in this section some techniques giving higher-order accuracy in space. The two main ideas are reducing the graph viscosity and introducing the consistent mass matrix. In particular, using the consistent mass matrix helps taming the dispersion errors.

83.1.1 Diminishing the graph viscosity

The high-order graph viscosity is denoted by $d_{ij}^{\mathrm{H},n}$, and the low-order graph viscosity defined in (81.19) is denoted by $d_{ij}^{\mathrm{L},n}$, where $n \geq 0$ is the time index. Thus, we consider the following low-order and high-order updates (see (81.18)):

$$m_i\big(\mathbf{U}_i^{\mathrm{L},n+1}-\mathbf{U}_i^n\big)+\sum_{j\in\mathcal{I}(i)}\tau\Big(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}-d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\Big)=\mathbf{0}, \tag{83.1a}$$

$$m_i\big(\mathbf{U}_i^{\mathrm{H},n+1}-\mathbf{U}_i^n\big)+\sum_{j\in\mathcal{I}(i)}\tau\Big(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}-d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n)\Big)=\mathbf{0}. \tag{83.1b}$$

Owing to Theorem 81.14, the low-order scheme (83.1a) is invariant domain preserving under the CFL condition

$$\max_{i\in\mathcal{A}_h^\circ}\gamma_i^n\le 1,\qquad \gamma_i^n:=\frac{2\tau|d_{ii}^{\mathrm{L},n}|}{m_i},\qquad d_{ii}^{\mathrm{L},n}:=\sum_{j\in\mathcal{I}(i)\backslash\{i\}}d_{ij}^{\mathrm{L},n}. \tag{83.2}$$

The first possibility to define the high-order update is to set the graph viscosity to zero, i.e., $d_{ij}^{\mathrm{H},n}=0$ for all $i,j\in\mathcal{A}_h$. The time-stepping scheme then becomes

$$\mathbf{U}_i^{\mathrm{G},n+1}:=\mathbf{U}_i^n-\frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)}\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}. \tag{83.3}$$

The superscript means that, up to the lumping of the mass matrix, $\mathbf{U}_i^{\mathrm{G},n+1}$ is nothing but the plain Galerkin approximation at the next time level. This choice may be suitable for linear problems, but it is often disastrous for nonlinear conservation equations, since, although limiting eventually makes the method invariant domain preserving, the resulting approximation may converge to an entropy-violating weak solution. In other words, once the Galerkin approximation goes in the wrong direction, limiting cannot put it back on the right track. A counterexample is constructed in Exercise 83.2; see also Remark 83.13. The reader should bear in mind that being invariant domain preserving and being entropy satisfying are not equivalent notions.

Smoothness-based graph viscosity. A better idea is to construct a high-order graph viscosity by proceeding as in §82.2.2. One difference between scalar equations and hyperbolic systems is that one needs now to choose the scalar quantity on which the smoothness indicator is based. One possibility is to choose a scalar-valued function $g:\mathcal{A}\to\mathbb{R}$ which could be an entropy or some scalar quantity associated with the problem in question. For the shallow-water equations, one could take the water height. For the Euler equations, one could take $g(\boldsymbol{u}):=\rho$ or $g(\boldsymbol{u}):=p$ (the density or the pressure) or g could be one of the generalized entropies $g(\boldsymbol{u}):=\rho f(\Phi(\boldsymbol{u}))$ (see Example 80.10). It is in general a good idea to choose an entropy since numerical experiments indicate that making the graph viscosity depend on the smoothness of an entropy may help the algorithm converge to an entropy satisfying solution. (See [169] for numerical experiments with the compressible Euler equations.) Once g is chosen, we set $\mathsf{G}_i^n:=g(\mathbf{U}_i^n)$ for all $i\in\mathcal{A}_h$, and

$$\alpha_i^n:=\frac{\left|\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\beta_{ij}(\mathsf{G}_j^n-\mathsf{G}_i^n)\right|}{\sum_{j\in\mathcal{I}(i)\backslash\{i\}}|\beta_{ij}(\mathsf{G}_j^n-\mathsf{G}_i^n)|}, \tag{83.4}$$

where the coefficients β_{ij} are meant to make the method linearity preserving. Since here it may not be relevant to be maximum principle preserving on $g(\boldsymbol{u}_h)$, one can take $\beta_{ij}:=-\int_D\nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x$ (i.e., it is not required that $\beta_{ij}\ge 0$). One can also use one of the linearity preserving techniques described in §82.2.2. Let $\psi\in\mathrm{Lip}([0,1];[0,1])$ be s.t. $\psi(1)=1$ and $\psi(0)=0$. The high-order smoothness-based graph viscosity is defined by setting

$$d_{ij}^{\mathrm{H},n}:=d_{ij}^{\mathrm{L},n}\max(\psi(\alpha_i^n),\psi(\alpha_j^n)),\qquad d_{ii}^{\mathrm{H},n}:=-\sum_{j\in\mathcal{I}(i)\backslash\{i\}}d_{ij}^{\mathrm{H},n}. \tag{83.5}$$

As discussed in Remark 82.14, this method produces an update $\boldsymbol{u}_h^{\mathrm{H},n+1}:=\sum_{i\in\mathcal{A}_h}\mathsf{U}_i^{\mathrm{H},n+1}\varphi_i$ that is (informally) second-order accurate in space in the $\boldsymbol{L}^1$-norm. The reader is referred to Remark 82.12 for some warnings concerning the choice of the function ψ.

Entropy-based graph viscosity. We now introduce a graph viscosity that is (informally) high-order for every polynomial degree, contrary to the one based on a smoothness indicator. It is also entropy consistent and close to being invariant domain preserving. We do not want to rely on the yet to be explained limiting process to enforce entropy consistency. We refer the reader to Lemma 3.2, Lemma 4.6, and §6.1 in [158] and §5.1 in [157] for counterexamples of methods that are invariant domain preserving but entropy violating (see also Exercise 81.4 and Exercise 83.2). Following an idea from Guermond et al. [166, 167], a high-order graph viscosity that is entropy consistent can be constructed by estimating a nondimensional entropy residual.

Given the current approximation $\mathbf{U}_i^n$, we first compute the Galerkin prediction $\mathbf{U}_i^{\mathrm{G},n+1}$ defined in (83.3). Let $(\eta(\boldsymbol{v}),\boldsymbol{F}(\boldsymbol{v}))$ be an entropy pair for the hyperbolic system (80.2). We estimate the entropy residual for the degree of freedom i by computing

$$\frac{m_i}{\tau}(\mathbf{U}_i^{\mathrm{G},n+1}-\mathbf{U}_i^n)\cdot\eta'(\mathbf{U}_i^n)+\sum_{j\in\mathcal{I}(i)}\boldsymbol{F}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}.$$

But, using the definition of $\mathbf{U}_i^{\mathrm{G},n+1}$, this is equivalent to computing

$$\sum_{j\in\mathcal{I}(i)}(\boldsymbol{F}(\mathbf{U}_j^n)-\eta'(\mathbf{U}_i^n)^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}.$$

This argument leads us to set

$$\begin{cases} N_i^n := \displaystyle\sum_{j\in\mathcal{I}(i)}(\boldsymbol{F}(\mathbf{U}_j^n)-(\eta'(\mathbf{U}_i^n))^{\mathsf{T}}\mathbb{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij},\\ D_i^n := \Big|\displaystyle\sum_{j\in\mathcal{I}(i)}\boldsymbol{F}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\Big|+\sum_{k\in\{1:m\}}\Big|\partial_{u_k}\eta(\mathbf{U}_i^n)\Big|\times\Big|\sum_{j\in\mathcal{I}(i)}\mathbb{f}_{u_k}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij}\Big|, \end{cases} \tag{83.6}$$

where $(\mathbb{f}_{u_k})_{k\in\{1:m\}}$ are the $\mathbb{R}^d$-valued rows of the flux $\mathbb{f}$. We then construct a normalized entropy viscosity ratio:

$$R_i^n := \frac{|\boldsymbol{N}_i^n|}{D_i^n}. \tag{83.7}$$

Notice that $R_i^n\in[0,1]$. Moreover, $N_i^n=0$ in the hypothetical case where $\eta:\mathbb{R}^m\to\mathbb{R}$ is linear. Finally, the high-order graph entropy viscosity is defined by setting

$$d_{ij}^{\mathrm{H},n} := d_{ij}^{\mathrm{L},n}\max(|R_i^n|,|R_j^n|),\qquad d_{ii}^{\mathrm{H},n} := -\sum_{i\neq j\in\mathcal{I}(i)}d_{ij}^{\mathrm{H},n}. \tag{83.8}$$

Remark 83.1 (Decay rate on R_i^n). Let us convince ourselves that R_i^n is at least one order smaller (in terms of the meshsize) than $d_{ij}^{\mathrm{L},n}$. Let us denote by $F''_{\max}$ and $f''_{\max}$ the maximum over $B_i^n:=\mathrm{conv}(\mathbf{U}_j^n)_{j\in\mathcal{I}(i)}$ of the matrix norm (induced by the Euclidean norm in $\mathbb{R}^m$) of the Hessian matrices $D^2\boldsymbol{F}$ and $D^2\mathbb{f}$, respectively. Recalling that $D\boldsymbol{F}(\mathbf{U})=\eta'(\mathbf{U})^{\mathsf{T}}D\mathbb{f}(\mathbf{U})$, we have

$$\begin{aligned}|N_i^n| &= \Big|\sum_{j\in\mathcal{I}(i)}\big(\boldsymbol{F}(\mathbf{U}_j^n)-\boldsymbol{F}(\mathbf{U}_i^n)-\eta'(\mathbf{U}_i^n)^{\mathsf{T}}(\mathbb{f}(\mathbf{U}_j^n)-\mathbb{f}(\mathbf{U}_i^n))\big)\cdot\boldsymbol{c}_{ij}\Big|\\ &\le \frac{1}{2}(F''_{\max}+\|\eta'(\mathbf{U}_i^n)\|_{\ell^2}f''_{\max})\max_{j\in\mathcal{I}(i)}\|\boldsymbol{c}_{ij}\|_{\ell^2}\sum_{j\in\mathcal{I}(i)}\|\mathbf{U}_j^n-\mathbf{U}_i^n\|_{\ell^2}^2.\end{aligned}$$

Similarly for the denominator, letting

$$F'_{\max}:=\max_{\boldsymbol{v}\in B_i^n}\|D\boldsymbol{F}(\boldsymbol{v})\|_{\ell^2(\mathbb{R}^{d\times m})},\qquad f'_{\max}:=\max_{\boldsymbol{v}\in B_i^n}\|D\mathbb{f}(\boldsymbol{v})\|_{\ell^2(\mathbb{R}^{md\times m})},$$

we have

$$|D_i^n| \le (F'_{\max} + \|\eta'(\mathbf{U}_i^n)\|_{\ell^2} f'_{\max}) \max_{j\in\mathcal{I}(i)} \|\boldsymbol{c}_{ij}\|_{\ell^2} \sum_{j\in\mathcal{I}(i)} \|\mathbf{U}_j^n - \mathbf{U}_i^n\|_{\ell^2}.$$

Hence, R_i^n scales at most like $\mathcal{O}(h/\ell_D)$, where h is the meshsize and ℓ_D a characteristic length of D, e.g., $\ell_D := \operatorname{diam}(D)$. □

Remark 83.2 (Euler equations, relative entropy). Letting β be an arbitrary constant, the change of entropy $\rho f(s) \to \rho(f(s)-\beta)$ for the Euler equations does not change the value of N_i^n. To account for this invariance, it is better to use the relative entropy $\eta_i^n(\boldsymbol{U}) = \rho^n(f(\Phi(\boldsymbol{U}))-f(\Phi(\boldsymbol{U}_i^n)))$ instead of $\rho^n f(\Phi(\boldsymbol{U})$, since this makes the definition of R_i^n invariant under the transformation $\rho f(s) \to \rho(\alpha f(s) - \beta)$ for all $\alpha, \beta \in \mathbb{R}$. □

83.1.2 Dispersion correction: consistent mass matrix

Recall that the two time-stepping schemes in (83.1) assume that the mass matrix is lumped. As emphasized in Remark 81.1, lumping is essential for these algorithms to be invariant domain preserving under a CFL condition. Although lumping the mass matrix does not affect the overall accuracy of the low-order method for smooth solutions, it nevertheless induces dispersion errors that have adverse effects when solving equations with nonsmooth initial data or with discontinuous solutions. It also impacts the accuracy of the higher-order methods. Some of these problems can be solved by using the consistent mass matrix. In particular, the dispersion phenomenon is well illustrated in dimension one on a uniform grid.

Proposition 83.3 (Dispersion error). *Consider a uniform mesh of size h over the interval $D := (0,1)$. Let $\{x_i\}_{i\in\mathcal{A}_h}$ be the mesh vertices. Let $(m_{ij})_{i\in\mathcal{A}_h, j\in\mathcal{I}(i)}$ be the coefficients of the consistent mass matrix for continuous $\mathbb{P}_1$ finite elements. Let $\beta \in \mathbb{R}$. Let $u \in C^6(D\times\mathbb{R}_+)$ solve $\partial_t u + \beta\partial_x u = 0$. The following holds true for all $i \in \mathcal{A}_h^\circ$:*

$$\partial_t u(x_i,t) + \beta\frac{u(x_{i+1},t) - u(x_{i-1},t)}{2h} = \beta h^2 C_u(x_i,t) + \mathcal{O}(h^4),$$
$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{m_i}\partial_t u(x_j,t) + \beta\frac{u(x_{i+1},t) - u(x_{i-1},t)}{2h} = -\beta h^4 \tilde{C}_u(x_i,t) + \mathcal{O}(h^6),$$

with $C_u(x_i,t) := \frac{1}{2}\partial_{xxx}u(x_i,t)$ and $\tilde{C}_u(x_i,t) := -\frac{1}{180}\partial_{xxxxx}u(x_i,t)$.

Proof. See Exercise 83.1. □

The above result shows that the leading term of the consistency error at the interior grid points is $\mathcal{O}(h^4)$ when using the consistent mass matrix, whereas it is $\mathcal{O}(h^2)$ when using the lumped mass matrix. In other words, the $\mathbb{P}_1$ approximation is superconvergent at the grid points when using the consistent mass matrix. The reader is referred to Christon et al. [85], Guermond and Pasquetti [154], Ainsworth [6], Thompson [274] for more details on this topic. The beneficial effects of the consistent mass matrix are particularly visible when solving problems with nonsmooth solutions.

In the rest of this chapter, we are going to assume that the provisional higher-order update $\mathbf{U}^{\mathrm{H},n+1}$ is computed using the consistent mass matrix, i.e., we replace (83.1b) by

$$\sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{\tau}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n) + \sum_{j\in\mathcal{I}(i)} \Big(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)\Big) = \mathbf{0}, \tag{83.9}$$

for some high-order graph viscosity $d_{ij}^{\mathrm{H},n}$ (its specific value is irrelevant in what follows). Let us emphasize that the price to pay to partially eliminate the dispersion errors is the loss of the

invariant domain property. This question is addressed in §83.2 using limiting techniques to post-process $\mathbf{U}^{\mathrm{H},n+1}$.

Remark 83.4 (Approximate inverse of the mass matrix). It is possible to avoid inverting the consistent mass matrix by proceeding for instance as in Abgrall [1], Guermond and Pasquetti [154]. Let $\mathcal{M}$ be the consistent mass matrix, $\overline{\mathcal{M}}$ the lumped mass matrix, and $\mathcal{B} := (\overline{\mathcal{M}}-\mathcal{M})\overline{\mathcal{M}}^{-1}$. It is shown in [154, Prop. 3.2] that one can approximate $\mathcal{M}^{-1} = \overline{\mathcal{M}}^{-1}(\mathcal{I}-\mathcal{B})^{-1}$ by $\overline{\mathcal{M}}^{-1}(\mathcal{I}+\mathcal{B})$ for $\mathbb{P}_1$ finite elements, since in this case the spectral radius of $\mathcal{B}$ is smaller than one; see Exercise 28.9. Denoting by $\mathbf{G}^n$ the vector with entries $\sum_{j\in\mathcal{I}(i)}(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{H},n}(\mathbf{U}_j^n-\mathbf{U}_i^n))$, the provisional higher-order update can be obtained as $\mathbf{U}^{\mathrm{H},n+1} = \mathbf{U}^n - \tau\overline{\mathcal{M}}^{-1}(\mathcal{I}+\mathcal{B})\mathbf{G}^n$, which only requires the inversion of the lumped mass matrix. □

83.2 Limiting

We show in this section how the provisional update $\mathbf{U}^{\mathrm{H},n+1}$ can be post-processed to make the map $\mathbf{U}^n \mapsto \mathbf{U}^{n+1}$ invariant domain preserving.

83.2.1 Key principles

The main idea for limiting consists of working with a low-order update $\mathbf{U}^{\mathrm{L},n+1}$, which is invariant domain preserving under the CFL condition (83.2), and a provisional higher-order update $\mathbf{U}^{\mathrm{H},n+1}$, which may step out of the invariant domain. The low-order update $\mathbf{U}^{\mathrm{L},n+1}$ is obtained from (83.1a), and the higher-order update $\mathbf{U}^{\mathrm{H},n+1}$ is obtained from (83.9). (The present techniques can also be adapted if $\mathbf{U}^{\mathrm{H},n+1}$ is obtained from (83.1b) using the lumped mass matrix.) An invariant domain preserving update $\mathbf{U}^{n+1}$ is then obtained by combining $\mathbf{U}^{\mathrm{L},n+1}$ and $\mathbf{U}^{\mathrm{H},n+1}$. Let us proceed naively for the time being, and let us assume that the invariant domain property consists of satisfying some constraint $\Psi(\mathbf{U}_i) \geq 0$ for all $i \in \mathcal{A}_h$. Here, a key property of the function Ψ is quasiconcavity.

Definition 83.5 (Quasiconcavity). *Given a convex set $B \subset \mathbb{R}^m$, we say that a continuous function $\Psi : B \to \mathbb{R}$ is* quasiconcave *if every* upper *level set of Ψ is* convex*, i.e., the set $L_\lambda(\Psi) := \{\mathbf{U} \in B \mid \Psi(\mathbf{U}) \geq \lambda\}$ is convex for all $\lambda \in \mathbb{R}$.*

Note that concavity implies quasiconcavity, but the converse is not true. For instance, the Gaussian function $\psi(x) := e^{-x^2}$ with $B := \mathbb{R}$ is quasiconcave but not concave. A simple result highlighting the difference between quasiconcavity and concavity is that a function $\Psi : B \to \mathbb{R}$ is quasiconcave iff for every finite set $\{\mathbf{U}_i\}_{i\in I} \subset B$ and all numbers $\{\theta_i\}_{i\in I} \subset [0,1]$ with $\sum_{i\in I}\theta_i = 1$, one has $\Psi(\sum_{i\in I}\theta_i\mathbf{U}_i) \geq \min_{i\in I}\Psi(\mathbf{U}_i)$ (see Exercise 83.3(i)). Notice that $\min_{i\in I}\Psi(\mathbf{U}_i)$ is smaller than $\sum_{i\in I}\theta_i\Psi(\mathbf{U}_i)$ which is the lower bound that is attained if Ψ is concave.

Example 83.6 (Euler equations). It is shown in Exercise 83.3(ii) that in the context of the Euler equations, the specific internal energy $e(\boldsymbol{u})$ and the specific entropy $\Phi(\boldsymbol{u})$ are quasiconcave functions. □

Let $i \in \mathcal{A}_h^\circ$ and assume that $\mathbf{U}_i^{\mathrm{L},n+1}$ is in the invariant domain $L_0(\Psi)$, i.e., $\Psi(\mathbf{U}_i^{\mathrm{L},n+1}) \geq 0$. The set $J_i^{n+1} := \{\ell \in [0,1] \mid \Psi((1-\ell)\mathbf{U}_i^{\mathrm{L},n+1} + \ell\mathbf{U}_i^{\mathrm{H},n+1}) \geq 0\}$ is nonempty (since $0 \in J_i^{n+1}$), so that it makes sense to define $\ell_i^{n+1} := \max_{\ell\in J_i^{n+1}}\ell$. Setting

$$\mathbf{U}_i^{n+1} := (1-\ell_i^{n+1})\mathbf{U}_i^{\mathrm{L},n+1} + \ell_i^{n+1}\mathbf{U}_i^{\mathrm{H},n+1} \tag{83.10}$$

then leads to $\Psi(\mathbf{U}_i^{n+1}) \geq 0$ for all $i \in \mathcal{A}_h^\circ$, i.e., $\mathbf{U}_i^{n+1}$ also lies in the invariant domain $L_0(\Psi)$. If the high-order graph viscosity is reasonably defined, one can reasonably expect that the above algorithm returns $\ell_i^{n+1} \approx 1$ most of the times, which would mean that $\mathbf{U}^{n+1}$ is very close to $\mathbf{U}^{\mathrm{H},n+1}$, i.e., it is reasonable to expect that $\mathbf{U}^{n+1}$ is high-order accurate.

We stop at this point to realize that the above program has one important flaw: it is not (globally) conservative. More precisely, using that $\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^n$ under appropriate boundary conditions (see Remark 81.4), we have

$$\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^n + \sum_{i\in\mathcal{A}_h} m_i\ell_i^{n+1}(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^{\mathrm{L},n+1}),$$

but we cannot conclude that $\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^n$ since there is no reason for the quantity $\sum_{i\in\mathcal{A}_h} m_i\ell_i^{n+1}(\mathbf{U}_i^{\mathrm{H},n+1}-\mathbf{U}_i^{\mathrm{L},n+1})$ to be zero. The rest of this section consists of addressing this issue. We are going to adapt the limiting technique described above to make it (globally) conservative. The two key words we are going to invoke from now on will be quasiconcavity and conservation.

83.2.2 Conservative algebraic formulation

In this section, we formulate a relation between $\mathbf{U}^{\mathrm{L},n+1}$ and $\mathbf{U}^{\mathrm{H},n+1}$ that properly accounts for (global) conservation. Since the (global) conservativity of the low-order and the high-order schemes implies that $\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{\mathrm{L},n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^n$ and $\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{\mathrm{H},n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^n$, we have

$$\sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{\mathrm{H},n+1} = \sum_{i\in\mathcal{A}_h} m_i\mathbf{U}_i^{\mathrm{L},n+1}. \tag{83.11}$$

Subtracting (83.1a) from (83.9), we obtain

$$m_i(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^{\mathrm{L},n+1}) = \sum_{j\in\mathcal{I}(i)} \Delta_{ij}(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n) + \tau(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n),$$

with $\Delta_{ij} := m_i\delta_{ij} - m_{ij}$. The above identity can be rewritten in a more concise way as follows:

$$m_i(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^{\mathrm{L},n+1}) = \sum_{j\in\mathcal{I}(i)} \mathbf{A}_{ij}^n, \tag{83.12}$$

with

$$\boldsymbol{A}_{ij}^n := \Delta_{ij}\big(\mathbf{U}_j^{\mathrm{H},n+1} - \mathbf{U}_j^n - (\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^n)\big) + \tau(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n), \tag{83.13}$$

where we used that $\sum_{j\in\mathcal{I}(i)} \Delta_{ij} = 0$. The key observation at this point is that the matrix $\mathbf{A}_{ij}^n$ is skew-symmetric. Then the (global) conservation property (83.11) can be proved (again) from (83.12) by simply summing (83.12) over $i \in \mathcal{A}_h$ and using the skew-symmetry of $\mathbf{A}_{ij}^n$.

From now on, irrespective of the exact way the provisional high-order update is computed, we assume that $\mathbf{U}_i^{\mathrm{H},n+1}$ and $\mathbf{U}_i^{\mathrm{L},n+1}$ satisfy (83.12) with the requirement that the coefficients $\mathbf{A}_{ij}^n \in \mathbb{R}^m$ are skew-symmetric. Since it is not guaranteed that $\mathbf{U}_i^{\mathrm{H},n+1}$ is in the invariant domain for all $i \in \mathcal{A}_h^\circ$, we are going to post-process $\mathbf{U}_i^{\mathrm{H},n+1}$. But instead of setting $\mathbf{U}_i^{n+1} := \mathbf{U}_i^{\mathrm{L},n+1} + \ell_i^{n+1}(\mathbf{U}^{\mathrm{H},n+1} - \mathbf{U}^{\mathrm{L},n+1})$ as we naively did in §83.2.1, we now set

$$m_i\mathbf{U}_i^{n+1} := m_i\mathbf{U}_i^{\mathrm{L},n+1} + \sum_{j\in\mathcal{I}(i)} \ell_{ij}^n\mathbf{A}_{ij}^n, \tag{83.14}$$

where the limiting coefficients ℓ_{ij}^n are going to be chosen in the interval $[0,1]$ with the symmetry constraint $\ell_{ij}^n = \ell_{ji}^n$ for all $j \in \mathcal{I}(i)$ and all $i \in \mathcal{A}_h$. Thus, the limiting coefficients are no longer attached to nodes but to pairs of nodes.

Lemma 83.7 (Conservation). *Assume that $\ell_{ij}^n = \ell_{ji}^n$ for all $j \in \mathcal{I}(i)$ and all $i \in \mathcal{A}_h$. Then $\sum_{i\in\mathcal{A}_h} m_i \mathbf{U}_i^{n+1} = \sum_{i\in\mathcal{A}_h} m_i \mathbf{U}_i^{\mathrm{L},n+1}$.*

Proof. Let $\mathcal{Z}_h := \{(i,j) \in \mathcal{A}_h \times \mathcal{A}_h \mid i \in \mathcal{I}(j), j \in \mathcal{I}(i)\}$. The symmetry of ℓ_{ij}^n and the skew-symmetry of $\mathbf{A}_{ij}^n$ imply that

$$\begin{aligned}\sum_{i\in\mathcal{A}_h} m_i\Big(\mathbf{U}_i^{n+1} - \mathbf{U}_i^{\mathrm{L},n+1}\Big) &= \sum_{i\in\mathcal{A}_h, j\in\mathcal{I}(i)} \ell_{ij}^n \mathbf{A}_{ij}^n \\ &= \sum_{(i,j)\in\mathcal{Z}_h} \frac{1}{2}\left(\ell_{ij}^n \mathbf{A}_{ij}^n + \ell_{ji}^n \mathbf{A}_{ji}^n\right) = \sum_{(i,j)\in\mathcal{Z}_h} \frac{1}{2}(\ell_{ij}^n \mathbf{A}_{ij}^n - \ell_{ij}^n \mathbf{A}_{ij}^n) = 0. \qquad \square\end{aligned}$$

Remark 83.8 (Anti-diffusion). Assume that the provisional high-order update $\mathbf{U}^{\mathrm{H},n+1}$ is computed with the lumped matrix instead of the consistent mass matrix and $d_{ij}^{\mathrm{H},n} = 0$, i.e., we use (83.3) and set $\mathbf{U}^{\mathrm{H},n+1} := \mathbf{U}^{\mathrm{G},n+1}$. Then $\mathbf{A}_{ij} = -\tau d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)$, and inserting the definition of $\mathbf{U}_i^{\mathrm{L},n+1}$ into (83.14) gives

$$m_i(\mathbf{U}_i^{n+1} - \mathbf{U}_i^n) + \sum_{j\in\mathcal{I}(i)} \tau\Big(\mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - (1-\ell_{ij}^n) d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n)\Big) = \mathbf{0}.$$

Hence, in this case, limiting is equivalent to replacing $d_{ij}^{\mathrm{L},n}$ by $(1-\ell_{ij}^n)d_{ij}^{\mathrm{L},n}$. In other words, limiting has an anti-diffusive effect, i.e., it reduces the graph viscosity. $\square$

Remark 83.9 (Approximate inverse of the mass matrix). Recalling Remark 83.4, one can avoid the inverse of the consistent mass matrix. Recalling that $\mathcal{B} := (\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{M}}^{-1}$, we observe that $\sum_{j\in\mathcal{I}(i)} \mathcal{B}_{ij} = 1 - \sum_{j\in\mathcal{I}(i)} \frac{m_{ij}}{m_j} = 0$ for all $i \in \mathcal{A}_h$. Subtracting the low-order equation (81.18) from the high-order equation $\mathbf{U}^{\mathrm{H},n+1} = \mathbf{U}^n - \tau\overline{\mathcal{M}}^{-1}(\mathcal{I}+\mathcal{B})\mathbf{G}^n$ obtained in Remark 83.4, we obtain

$$\begin{aligned}m_i(\mathbf{U}_i^{\mathrm{H},n+1} - \mathbf{U}_i^{\mathrm{L},n+1}) &= \tau \sum_{j\in\mathcal{I}(i)} -(\delta_{ij} + \mathcal{B}_{ij})\mathbf{G}_j^n + \mathbb{f}(\mathbf{U}_j^n)\cdot\boldsymbol{c}_{ij} - d_{ij}^{\mathrm{L},n}(\mathbf{U}_j^n - \mathbf{U}_i^n) \\ &= \tau \sum_{j\in\mathcal{I}(i)} -\mathcal{B}_{ij}\mathbf{G}_j^n + (d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n).\end{aligned}$$

Setting $\mathbf{A}_{ij}^n := -\tau\mathcal{B}_{ij}(\mathbf{G}_j^n - \mathbf{G}_i^n) + \tau(d_{ij}^{\mathrm{H},n} - d_{ij}^{\mathrm{L},n})(\mathbf{U}_j^n - \mathbf{U}_i^n)$, which is legitimate since $\sum_{j\in\mathcal{I}(i)} \mathcal{B}_{ij} = 0$, the above identity takes the same form as (83.12). $\square$

83.2.3 Boris–Book–Zalesak's limiting for scalar equations

In this section, we introduce a limiting technique developed by Boris and Book [39] and Zalesak [291] for scalar equations and called *Flux Corrected Transport (FCT)*. We refer the reader to Kuzmin et al. [209] for a review of this topic.

We drop the time index n whenever the context is unambiguous. Let us assume that the low-order update satisfies some minimum and maximum principle, say, there are two vectors $\mathsf{U}^{\min} \in \mathbb{R}^I$ and $\mathsf{U}^{\max} \in \mathbb{R}^I$ s.t.

$$\mathsf{U}_i^{\mathrm{L}} \in [\mathsf{U}_i^{\min}, \mathsf{U}_i^{\max}], \qquad \forall i \in \mathcal{A}_h. \tag{83.15}$$

For instance, $\mathsf{U}_i^{\min} := \min_{j\in\mathcal{I}(i)} \mathsf{U}_i^n$ and $\mathsf{U}_i^{\max} := \max_{j\in\mathcal{I}(i)} \mathsf{U}_i^n$ are possible definitions of $\mathsf{U}_i^{\min}$ and $\mathsf{U}_i^{\max}$.

There are (infinitely) many ways to define the limiting coefficients ℓ_{ij}. The method described in [291] consists of first computing the following coefficients P_i^+, P_i^-, Q_i^+, Q_i^-, R_i^+, and R_i^- for all $i \in \mathcal{A}_h$:

$$P_i^+ := \sum_{j\in\mathcal{I}(i)} \max\{0, \mathsf{A}_{ij}\}, \qquad P_i^- := \sum_{j\in\mathcal{I}(i)} \min\{0, \mathsf{A}_{ij}\}, \tag{83.16}$$

$$Q_i^+ := m_i(\mathsf{U}_i^{\max} - \mathsf{U}_i^{\mathrm{L}}), \qquad Q_i^- := m_i(\mathsf{U}_i^{\min} - \mathsf{U}_i^{\mathrm{L}}), \tag{83.17}$$

$$R_i^+ := \begin{cases} \min\{1, \frac{Q_i^+}{P_i^+}\} & P_i^+ \neq 0, \\ 1 & P_i^+ = 0, \end{cases} \qquad R_i^- := \begin{cases} \min\{1, \frac{Q_i^-}{P_i^-}\} & P_i^- \neq 0, \\ 1 & P_i^- = 0. \end{cases} \tag{83.18}$$

Then the limiting coefficients ℓ_{ij} are defined as follows:

$$\ell_{ij} := \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } \mathsf{A}_{ij} \geq 0, \\ \min\{R_i^-, R_j^+\} & \text{otherwise.} \end{cases} \tag{83.19}$$

Lemma 83.10 (Limiting coefficients). *The definitions* (83.16)–(83.19) *imply* $P_i^- \leq 0 \leq P_i^+$, $Q_i^- \leq 0 \leq Q_i^+$, $0 \leq R_i^-$, $0 \leq R_i^+$ *for all* $i \in \mathcal{A}_h$, *and*

$$\ell_{ij} \in [0,1], \qquad \ell_{ij} = \ell_{ji}, \qquad \forall j \in \mathcal{I}(i),\ \forall i \in \mathcal{A}_h. \tag{83.20}$$

Proof. The properties on $P_i^\pm$, $Q_i^\pm$, $R_i^\pm$ follow immediately from the definitions (83.16)–(83.18) and the assumption (83.15). The definitions of R_i^+ and R_j^- imply that $0 \leq \ell_{ij} \leq 1$. Let us now prove that $\ell_{ij} = \ell_{ji}$. Assume that $\mathsf{A}_{ij} \geq 0$. Then $\mathsf{A}_{ji} = -\mathsf{A}_{ij} \leq 0$. The definitions of ℓ_{ij} and ℓ_{ji} imply in turn that $\ell_{ij} = \min\{R_i^+, R_j^-\}$ and $\ell_{ji} = \min\{R_j^-, R_i^+\}$, i.e., $\ell_{ij} = \ell_{ji}$. The proof for the case $\mathsf{A}_{ij} \leq 0$ is identical. □

Theorem 83.11 (Maximum principle). *Let* $\mathsf{U}^{\min}, \mathsf{U}^{\max} \in \mathbb{R}^I$ *be s.t.* (83.15) *holds true. Then the update given by* (83.14) *with* ℓ_{ij} *defined in* (83.19) *satisfies*

$$\mathsf{U}_i^{n+1} \in [\mathsf{U}_i^{\min}, \mathsf{U}_i^{\max}], \qquad \forall i \in \mathcal{A}_h^\circ. \tag{83.21}$$

Proof. Assume that $P_i^+ \neq 0$. By (83.14) and the definition of ℓ_{ij}, we have

$$\begin{aligned} m_i(\mathsf{U}_i^{n+1} - \mathsf{U}_i^{\mathrm{L}}) &= \sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathsf{A}_{ij} \leq \sum_{\substack{j\in\mathcal{I}(i)\\ 0\leq\mathsf{A}_{ij}}} \ell_{ij}\mathsf{A}_{ij} = \sum_{\substack{j\in\mathcal{I}(i)\\ 0\leq\mathsf{A}_{ij}}} \min\{R_i^+, R_j^-\}\mathsf{A}_{ij} \\ &\leq \sum_{\substack{j\in\mathcal{I}(i)\\ 0\leq\mathsf{A}_{ij}}} R_i^+\mathsf{A}_{ij} \leq \sum_{\substack{j\in\mathcal{I}(i)\\ 0\leq\mathsf{A}_{ij}}} \frac{Q_i^+}{P_i^+}\mathsf{A}_{ij} \\ &= \frac{Q_i^+}{P_i^+} \sum_{j\in\mathcal{I}(i)} \max\{0, \mathsf{A}_{ij}\} = Q_i^+ = m_i(\mathsf{U}_i^{\max} - \mathsf{U}_i^{\mathrm{L}}), \end{aligned}$$

which proves that $\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\max}$ when $P_i^+ \neq 0$. If $P_i^+ = 0$, then $m_i(\mathsf{U}_i^{n+1} - \mathsf{U}_i^{\mathrm{L}}) \leq 0 \leq m_i(\mathsf{U}_i^{\max} - \mathsf{U}_i^{\mathrm{L}})$, which proves again that $\mathsf{U}_i^{n+1} \leq \mathsf{U}_i^{\max}$. The lower bound, $\mathsf{U}_i^{\min} \leq \mathsf{U}_i^{n+1}$, is proved similarly. □

Remark 83.12 ($\mathsf{U}_i^{\max}$, $\mathsf{U}_i^{\min}$). The maximum principle is satisfied independently of the value of $\mathsf{U}_i^{\max}$ and $\mathsf{U}_i^{\min}$, provided that (83.15) holds true. □

Remark 83.13 (FCT counterexample). One must be careful when using limiting. For instance, without changing the low-order update, one could consider the method for which the provisional high-order update is the Galerkin solution, i.e., $d_{ij}^{\mathrm{H},n} = 0$. Then applying FCT to the pair low-order/Galerkin produces a method that is high-order accurate in space and maximum principle preserving. This recipe is indeed a good method for solving linear equations, but it may fail to converge to the entropy solution when solving nonlinear equations. See Exercise 83.2 and [158, Lem. 4.6] for a counterexample. □

83.2.4 Convex limiting for hyperbolic systems

We return in this section to hyperbolic systems. It is no longer possible to apply the FCT methodology because the maximum principle is no longer meaningful, even if the system is linear. To be fully convinced, consider the one-dimensional linear wave equation $\partial_t\rho+\rho_0\partial_x v = 0$, $\partial_t v+\frac{a^2}{\rho_0}\partial_x\rho = 0$, with constants $\rho_0 > 0$ and $a > 0$. In this case, one may wonder whether ρ, the scalar component of $\boldsymbol{u} = (\rho, v)^\mathsf{T}$, satisfies some sort of maximum principle. It is shown in Exercise 80.7 that it is not the case: one can always find initial data, $(\rho_0, v_0)^\mathsf{T}$, s.t. either $\min_{\{x,t\}}\rho(x,t) < \min_x\rho_0(x)$ or $\max_{\{x,t\}}\rho(x,t) > \max_x\rho_0(x)$. The situation is even worse in dimension three as shown in Exercise 80.6.

We have seen in §80.2.3 that the notion of maximum principle must be replaced by the notion of invariant set. But, this notion is not rich enough for our purpose since it is global. For the Euler equations for instance, one family of natural invariant sets is $B_r := \{\boldsymbol{u} := (\rho, \boldsymbol{m}, E)^\mathsf{T} \mid \rho > 0,\ e(\boldsymbol{u}) \geq 0,\ s(\rho, e(\boldsymbol{u})) \geq r\}$, $r \in \mathbb{R}$. But asserting that conditions like $\rho > 0$, $e(\boldsymbol{u}) \geq 0$, $s(\rho, e(\boldsymbol{u})) \geq r$ be satisfied for the update $\mathbf{U}^{n+1}$ is far poorer than enforcing a bound like (83.21) where the values of the lower and upper bounds are local. To be really efficient and to eliminate (or reduce) local "oscillations", limiting should be local. We now present a technique introduced in Guermond et al. [169, 171] and called *convex limiting* that does exactly that.

Let us recall that the intermediate states $\overline{\mathbf{U}}_{ij}^n$, for all $j \in \mathcal{I}(i)\backslash\{i\}$ and all $i \in \mathcal{A}_h^\circ$, defined by

$$\overline{\mathbf{U}}_{ij}^n := \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n) - (\mathbb{f}(\mathbf{U}_j^n) - \mathbb{f}(\mathbf{U}_i^n))\frac{\boldsymbol{c}_{ij}}{2d_{ij}^n} \tag{83.22}$$

are essential to establish the invariant domain property of the scheme (83.1a) under the CFL condition (83.2). In particular, setting $\overline{\mathbf{U}}_{ii}^n := \mathbf{U}_i^n$, we have $\mathbf{U}_i^{\mathrm{L},n+1} \in \mathrm{conv}_{j\in\mathcal{I}(i)}(\overline{\mathbf{U}}_{ij}^n)$, which immediately implies the following result.

Lemma 83.14 (Lower bound). *Let $\Psi : \mathcal{A} \to \mathbb{R}^m$ be an arbitrary quasiconcave function. Then under the CFL condition* (83.2)*, we have $\Psi(\mathbf{U}_i^{\mathrm{L},n+1}) \geq \min_{j\in\mathcal{I}(i)}\Psi(\overline{\mathbf{U}}_{ij}^n)$ for all $i \in \mathcal{A}_h^\circ$.*

We now have the right localizing tool in hand. Given a quasiconcave function Ψ, we are going to construct some limiting technique so that the post-processed update $\mathbf{U}_i^{n+1}$ satisfies $\Psi(\mathbf{U}_i^{n+1}) \geq \min_{j\in\mathcal{I}(i)}\Psi(\overline{\mathbf{U}}_{ij}^n)$ for all $i \in \mathcal{A}_h^\circ$.

Remark 83.15 (Oscillations). One should be careful about the meaning of the generic word "oscillations" when working with hyperbolic systems, since this concept is essentially scalar. It usually refers to the graph of some scalar-valued function that unexpectedly goes above or below some reference value and then comes back within the expected bounds. This notion is somewhat irrelevant for hyperbolic systems. For instance, there exist hyperbolic systems such that the invariant domain preserving technique (83.1a) produces an approximate solution with one Cartesian component that "oscillates", but the said approximation actually stays in every invariant set of the PDEs (see, e.g., [157, §5.3]). □

We now drop the time index n and assume that the higher-order provisional update $\mathbf{U}^{\mathrm{H}}$ and the low-order update $\mathbf{U}^{\mathrm{L}}$ are related by

$$m_i \mathbf{U}_i^{\mathrm{H}} = m_i \mathbf{U}_i^{\mathrm{L}} + \sum_{j \in \mathcal{I}(i)} \mathbf{A}_{ij},$$

where $\mathbf{A}_{ij} = -\mathbf{A}_{ji}$ (see (83.12) in §83.2.2). We now depart from the FCT algorithm by introducing parameters $\theta_{ij} \in (0,1)$ for all $j \in \mathcal{I}(i)$ and all $i \in \mathcal{A}_h$. Two typical examples are as follows: (1) $\theta_{ij} = \frac{m_{ij}}{m_i}$, $j \in \mathcal{I}(i)$; (2) $\theta_{ii} = 0$ and $\theta_{ij} = (\mathrm{card}(\mathcal{I}(i)) - 1)^{-1}$ for all $j \in \mathcal{I}(i) \backslash \{i\}$. (The numerical illustrations reported in [169] have been done with the second choice.) Note that both examples satisfy the important property

$$\sum_{j \in \mathcal{I}(i)} \theta_{ij} = 1, \qquad \forall i \in \mathcal{A}_h. \tag{83.23}$$

Then we have

$$\mathbf{U}_i^{\mathrm{H}} = \sum_{j \in \mathcal{I}(i)} \theta_{ij} (\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) \quad \text{with} \quad \mathbf{P}_{ij} := \frac{1}{m_i \theta_{ij}} \mathbf{A}_{ij}, \tag{83.24}$$

i.e., $\mathbf{U}_i^{\mathrm{H}}$ is a convex combination of $\{\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}\}_{j \in \mathcal{I}(i)}$. Our main result is the following.

Theorem 83.16 (Convex limiting). *Let $\Psi^0 : \mathcal{A} \to \mathbb{R}^m$ be a quasiconcave function and assume that $\Psi^0(\mathbf{U}_i^{\mathrm{L}}) \geq 0$. For all $i \in \mathcal{A}_h$ and all $j \in \mathcal{I}(i)$, let $\ell_j^i \in [0,1]$ be defined by*

$$\ell_j^i := \begin{cases} 1 & \text{if } \Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) \geq 0, \\ \max\{\ell \in [0,1] \mid \Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell \mathbf{P}_{ij}) \geq 0\} & \text{otherwise.} \end{cases}$$

(i) We have $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0, \ell_j^i]$. (ii) Setting $\ell_{ij} := \min(\ell_j^i, \ell_i^j)$, we have $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij}) \geq 0$ and $\ell_{ij} = \ell_{ji}$ for all $j \in \mathcal{I}(i)$ and all $i \in \mathcal{A}_h$. (iii) The following inequality holds true:

$$\Psi^0\Big(\sum_{j \in \mathcal{I}(i)} \theta_{ij} (\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij}) \Big) \geq 0.$$

Proof. Consider the upper level set $L_0(\Psi^0) := \{\mathbf{V} \in \mathcal{A} \mid \Psi^0(\mathbf{V}) \geq 0\}$ which is a convex set since Ψ^0 is quasiconcave.
(i) First, if $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) \geq 0$, we have $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell \mathbf{P}_{ij}) \geq 0$ for all $\ell \in [0,1]$, because $\mathbf{U}_i^{\mathrm{L}} \in L_0(\Psi^0)$, $\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij} \in L_0(\Psi^0)$, and $L_0(\Psi^0)$ is a convex set. Second, if $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) < 0$, we observe that ℓ_j^i is uniquely defined, and for all $\ell \in [0, \ell_j^i]$, we have $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell \mathbf{P}_{ij}) \geq 0$ by the same argument as above.
(ii) Since $\ell_{ij} = \min(\ell_j^i, \ell_i^j) \leq \ell_j^i$, the above construction implies that $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij}) \geq 0$. Moreover, the symmetry of ℓ_{ij} results from its definition.
(iii) All the limited states $\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij}$ are in $L_0(\Psi^0)$ for all $j \in \mathcal{I}(i)$. Since the set $L_0(\Psi^0)$ is convex, the convex combination $\sum_{j \in \mathcal{I}(i)} \theta_{ij} (\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij})$ is in $L_0(\Psi^0)$, i.e., $\Psi^0\big(\sum_{j \in \mathcal{I}(i)} \theta_{ij} (\mathbf{U}_i^{\mathrm{L}} + \ell_{ij} \mathbf{P}_{ij}) \big) \geq 0$. □

The idea behind Theorem 83.16 is illustrated in Figure 83.1. This theorem is used as follows. Given some quasiconcave function $\Psi : \mathcal{A} \to \mathbb{R}$, we define $\Psi_i^0(\mathbf{V}) := \Psi(\mathbf{V}) - \min_{j \in \mathcal{I}(i)} \Psi(\overline{\mathbf{U}}_{ij})$ for all $i \in \mathcal{A}_h^\circ$. Owing to Lemma 83.14, we have $\Psi_i^0(\mathbf{U}_i^{\mathrm{L}}) \geq 0$ (under the CFL condition (83.2)), which is the key assumption in Theorem 83.16. Then we compute the symmetric limiting matrix ℓ_{ij} as in the theorem, and we set

$$m_i \mathbf{U}_i^{n+1} := m_i \mathbf{U}_i^{\mathrm{L}} + \sum_{j \in \mathcal{I}(i)} \ell_{ij} \mathbf{A}_{ij}. \tag{83.25}$$

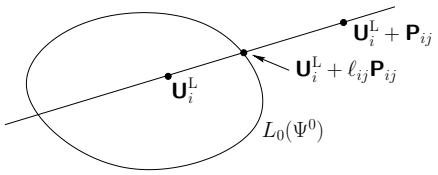

Figure 83.1: Convex limiting: illustration of Theorem 83.16.

This construction implies that $\Psi(\mathbf{U}_i^{n+1}) \geq \min_{j\in\mathcal{I}(i)} \Psi(\overline{\mathbf{U}}_{ij})$, which is the expected invariant domain property.

Remark 83.17 (Multiple limiting). In general, we have to consider families of quasiconcave functions $\{\{\Psi_i^l\}_{i\in\mathcal{V}}\}_{l\in\mathcal{L}}$, $\Psi_i^l : \mathcal{B}^l \to \mathbb{R}$, where $\mathcal{B}^l \subset \mathbb{R}^m$ is the convex admissible set for the function Ψ_i^l. The list $\mathcal{L}$ describes the nature of the functions. It is readily verified that the function $\Psi := \min_{l\in\mathcal{L}} \Psi_i^l : \bigcap_{l\in\mathcal{L}} \mathcal{B}^l \to \mathbb{R}$ is quasiconcave and that its upper level sets are such that $L_\Psi(\lambda) = \bigcap_{l\in\mathcal{L}} L_{\Psi_i^l}(\lambda)$. The list $\mathcal{L}$ is sometimes ordered in the sense that $\mathcal{B}^{l'} \subset \mathcal{B}^l$ if $l' \geq l$. Let us illustrate this concept with the compressible Euler equations. Usually, one starts with $B^1 := \mathbb{R}^m$ to enforce a local minimum principle on the density (which implies positivity of the density). One can also take $\mathcal{B}^2 := \mathbb{R}^m$ to enforce a local maximum principle on the density by using $\Psi(\mathbf{U}) := -\rho$. Then one can consider $B^3 := \{\mathbf{U} \in \mathcal{B}^1 \mid \rho > 0\}$ to enforce a local minimum principle on the (specific) internal energy (which implies positivity of the (specific) internal energy). One can finally set $B^4 := \{\mathbf{U} \in \mathcal{B}^2 \mid e(\mathbf{U}) > 0\}$ to enforce a local minimum principle on the specific entropy. □

Example 83.18 (Linear Ψ). It can happen that Ψ is linear. For the Euler equations for instance, the density ρ and its opposite $-\rho$ are linear functionals of the conserved variable $\boldsymbol{u} := (\rho, \boldsymbol{m}, E)^\mathsf{T}$. One can then apply Theorem 83.16 by setting either $\Psi(\boldsymbol{u}) := \rho - \rho_i^{\min}$ or $\Psi(\boldsymbol{u}) := \rho_i^{\max} - \rho$ with $\rho_i^{\min} := \min_{j\in\mathcal{I}(i)} \overline{\rho}_{ij}^n$ and $\rho_i^{\max} := \max_{j\in\mathcal{I}(i)} \overline{\rho}_{ij}^n$. Limiting w.r.t. these two functionals gives $\rho_i^{\min} \leq \rho_i^{n+1} \leq \rho_i^{\max}$. The computation of ℓ_j^i is trivial in this case. Provided the CFL number is small enough, Item (ii) in Theorem 81.14 implies that $\rho_i^{\min} := \min_{j\in\mathcal{I}(i)} \overline{\rho}_{ij}^n > 0$, i.e., $\rho_i^{n+1} > 0$ for all $i \in \mathcal{A}_h^\circ$. □

Example 83.19 (Quadratic Ψ). If Ψ is quadratic, computing the parameter ℓ_j^i defined in Theorem 83.16 amounts to solving a quadratic equation. After setting $a := \frac{1}{2}\mathbf{P}_{ij}^\mathsf{T} D^2\Psi \mathbf{P}_{ij}$, $b := (D\Psi(\mathbf{U}^{\mathrm{L}}))^\mathsf{T}\mathbf{P}_{ij}$, and $c := \Psi(\mathbf{U}^{\mathrm{L}})$, we have $\Psi(\mathbf{U}^{\mathrm{L}} + t\mathbf{P}_{ij}) = \frac{1}{2}at^2 + bt + c$. Let t_0 be the smallest positive root of the equation $at^2 + bt + c = 0$ with the convention that $t_0 := 1$ if the equation has no positive root. Then we set $\ell_j^i := \min(t_0, 1)$. □

Remark 83.20 (Reduction to a quadratic functional). Assume that one wants to limit the quasiconcave functional $\Psi : \mathcal{B} \to \mathbb{R}$. Assume that there exists $\phi : \mathcal{B} \to \mathbb{R}$ and $\ell^\phi \in (0, 1]$ s.t. $\phi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) > 0$ for all $\ell \in [0, \ell^\phi]$. Assume also that $[0, \ell^\phi] \ni \ell \mapsto (\phi\Psi)(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij})$ is quadratic, $\Psi(\mathbf{U}^{\mathrm{L}}) > 0$ and $\Psi(\mathbf{U}^{\mathrm{L}} + \ell^\phi\mathbf{P}_{ij}) < 0$. Then for all $\ell \in [0, \ell^\phi]$ we have $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) \geq 0$ iff $(\phi\Psi)(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) \geq 0$ and $\Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) = 0$ iff $(\phi\Psi)(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) = 0$. Hence, instead of doing a nonlinear line search on the quasiconcave functional $[0, \ell^\phi] \ni \ell \mapsto \Psi(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij})$, one can compute the limiter $\ell_j^i \in [0, \ell^\phi]$ defined in Theorem 83.16 by simply solving the quadratic equation $(\phi\Psi)(\mathbf{U}^{\mathrm{L}} + \ell\mathbf{P}_{ij}) = 0$, and this can be done as explained in Example 83.19. Whether the functional $\phi\Psi$ is quasiconcave or not is irrelevant here. □

Example 83.21 (Euler equations). Let us illustrate the technique from Remark 83.20 with the Euler equations. Consider the internal energy $\mathcal{E}(\boldsymbol{u}) := E - \frac{1}{2}\rho^{-1}\boldsymbol{m}^2$. This function is concave

because its second order Fréchet derivative at $\boldsymbol{u}$ is s.t. $D^2\mathcal{E}(\boldsymbol{u})((a,\boldsymbol{b},c),(a,\boldsymbol{b},c)) = -\frac{1}{\rho}(\frac{\boldsymbol{m}}{\rho}a-\boldsymbol{b})^2$ for all $(a,\boldsymbol{b},c)\in\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. Hence, the specific internal energy $e(\boldsymbol{u}) := \frac{1}{\rho}\mathcal{E}(\boldsymbol{u})$ is quasiconcave; see Exercise 83.3. Let us set $e_i^{\min} := \min_{j\in\mathcal{I}(i)} e(\overline{\mathbf{U}}_{ij}^n)$, $\Psi(\boldsymbol{u}) := e(\boldsymbol{u}) - e_i^{\min}$, and $\phi(\boldsymbol{u}) := \rho^2$. Notice that $(\phi\Psi)(\boldsymbol{u}) := \rho E - \frac{1}{2}\boldsymbol{m}^2 - \rho^2 e_i^{\min}$ is quadratic (but this functional is neither convex nor concave). Assume that limiting on the density is done first with the limiting parameter $\ell_j^{\rho,i}$. Then computing ℓ_j^i can be done easily by solving a quadratic equation as explained above, i.e., $\ell_j^i = \min(t_0, 1, \ell_j^{\rho,i})$. After limiting, we have $\rho_i^{n+1}E_i^{n+1} - \frac{1}{2}(\boldsymbol{m}_i^{n+1})^2 = (\rho_i^{n+1})^2 e(\mathbf{U}_i^{n+1}) \geq (\rho_i^{n+1})^2 e_i^{\min}$. Assuming that the CFL number is small enough so that Item (ii) in Theorem 81.14 implies $\rho_i^{\min} := \min_{j\in\mathcal{I}(i)} \overline{\rho}_{ij}^n > 0$ and $e_i^{\min} := \min_{j\in\mathcal{I}(i)} e(\overline{\mathbf{U}}_{ij}^n) > 0$, we obtain $e(\mathbf{U}_i^{n+1}) \geq e_i^{\min} > 0$ (because $\rho_i^{n+1} \geq \rho_i^{\min} > 0$). The above technique can also be used to limit with respect to the kinetic energy $\mathsf{k}(\boldsymbol{u}) := \frac{1}{2}\rho^{-1}\boldsymbol{m}^2$. The negative of the kinetic energy is quasiconcave since $\rho\mathsf{k} = -\frac{1}{2}\boldsymbol{m}^2$ is concave; see Exercise 83.3. Then setting $\mathsf{k}^{\max} := \max_{j\in\mathcal{I}(i)} \overline{\mathsf{k}}_{ij}^n$, limiting the quasiconcave functional $\Psi(\boldsymbol{u}) := -\mathsf{k} + \mathsf{k}^{\max}$ by using the quadratic functional $\rho\Psi(\boldsymbol{u}) := -\frac{1}{2}\boldsymbol{m}^2 + \rho\mathsf{k}^{\max}$ gives $\mathsf{k}(\mathbf{U}_i^{n+1}) \leq \mathsf{k}^{\max}$. □

In the general case, the computation of the limiting parameters ℓ_j^i can be done as follows. We observe that the equation $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij}) = 0$ has at most two roots (possibly equal) because the upper level set $L_0(\Psi^0) = \{\mathbf{U}\in\mathcal{A} \mid \Psi^0(\mathbf{U}) \geq 0\}$ is convex and every line that intersects $L_0(\Psi^0)$ crosses the boundary of $L_0(\Psi^0)$ at two points (at most), say $t_- \leq t_+$, ($t_- = t_+$ when the line is tangent to the boundary of $L_0(\Psi^0)$). Notice that $t_- \leq 0$ since $\Psi^0(\mathbf{U}_i^{\mathrm{L}}) \geq 0$. (i) If $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) \geq 0$, then $t_+ \geq 1$ and the entire segment $\{\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij} \mid t\in[0, t_0 = 1]\}$ is in $L_0(\Psi^0)$ by convexity. Thus, we set $t_0 := 1$. (ii) If $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) < 0$ and $\Psi^0(\mathbf{U}_i^{\mathrm{L}}) > 0$, then $t_+ \in (0,1)$, and setting $t_0 := t_+$, the entire segment $\{\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij} \mid t\in[0,t_0]\}$ is in $L_0(\Psi^0)$. Note that in this case t_+ is the unique positive root to the equation $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij}) = 0$. (iii) Assume finally that $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + \mathbf{P}_{ij}) < 0$ and $\Psi^0(\mathbf{U}_i^{\mathrm{L}}) = 0$. There are two possibilities: (iii.a) If $D\psi(\mathbf{U}_i^{\mathrm{L}})\cdot\mathbf{P}_{ij} \leq 0$ then by convexity $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij}) < 0$ for all $t > 0$. Hence, $t_+ = 0$ is the largest nonnegative root of the equation $\Psi^0(\mathbf{U}_i^{\mathrm{L}} + t\mathbf{P}_{ij}) = 0$ and therefore $t_0 = t_+ = 0$. (iii.b) In the other case, $D\psi(\mathbf{U}_i^{\mathrm{L}})\cdot\mathbf{P}_{ij} > 0$, we have that $0 < t_+ < 1$ and we set $t_0 := t_+$. In all the cases, the limiting coefficient is obtained by setting $\ell_j^i := t_0$.

Example 83.22 (Newton-secant algorithm). Let us illustrate the general situation on the Euler equations by using limiting to enforce the minimum principle on the specific entropy. Notice that Φ is quasiconcave since $\rho\Phi$ is concave; see Exercise 83.3. Let $\Phi_i^{\min} := \min_{j\in\mathcal{I}(i)} \Phi(\mathbf{U}_j^n)$ and set $\Psi(\mathbf{U}) := \rho\Phi(\mathbf{U}) - \rho\Phi_i^{\min}$. Since $\rho\Phi$ is concave, the function $h(t) := \Psi(\mathbf{U}_i^{\mathrm{L},n+1} + t\mathbf{P}_{ij})$ is concave and solving the equation $h(t) = 0$ can be done very efficiently. If $h(1) \geq 0$, we set $t_0 := 1$, and if $h(1) < 0$, we can combine the secant and Newton's method to find the unique root $t_0 \in [0,1]$ such that $h(t_0) = 0$. The main interest of the Newton-secant technique is that for every threshold ϵ, the algorithm is guaranteed to return an answer t_ϵ such that $h(t_\epsilon) \geq 0$ (whereas Newton's algorithm with $t := 1$ as initial guess always returns $h(t_\epsilon) \leq 0$ independently of the threshold). Other implementation details are reported in Guermond et al. [171, §7.5.4]. □

The limiting process described above can be iterated multiple times by observing from (83.25) that

$$\mathbf{U}_i^{\mathrm{H},n+1} = \mathbf{U}_i^{\mathrm{L},n+1} + \frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} \ell_{ij}\mathbf{A}_{ij}^n + \frac{1}{m_i}\sum_{j\in\mathcal{I}(i)} (1-\ell_{ij})\mathbf{A}_{ij}^n. \tag{83.26}$$

Then, by setting $\mathbf{U}^{(0)} := \mathbf{U}_i^{\mathrm{L},n+1}$ and $\mathbf{A}_{ij}^{(0)} = \mathbf{A}_{ij}^n$, one can iteratively repeat the limiting by proceeding as described in Algorithm 83.1. It is common to take $k_{\max} := 2$ since further iterations generally do not improve the accuracy of the approximation.

Algorithm 83.1 Iterative limiting

Require: $\mathsf{U}^{\mathrm{L},n+1}$, A_{ij}^n, $k_{\max}$
Ensure: U^{n+1}
  Set $\mathsf{U}^{(0)} := \mathsf{U}_i^{\mathrm{L},n+1}$ and $\mathsf{A}^{(0)} := \mathsf{A}^n$
  for $k = 0$ to $k_{\max} - 1$ **do**
    Compute limiter $\ell^{(k)}$
    Update $\mathsf{U}^{(k+1)} = \mathsf{U}^{(k)} + \frac{1}{m_i} \sum_{j \in \mathcal{I}(i)} \ell_{ij}^{(k)} \mathsf{A}_{ij}^{(k)}$
    Update $\mathsf{A}_{ij}^{(k+1)} = (1 - \ell_{ij}^{(k)}) \mathsf{A}_{ij}^{(k)}$
  end for
  $\mathsf{U}^{n+1} := \mathsf{U}^{(k_{\max})}$

Remark 83.23 (Bound relaxation). The limiting method described in this chapter suffers from the same convergence deficiencies as the viscosity reduction techniques described in §82.2 for scalar conservation equations (see Remark 82.14). It delivers second-order accuracy in space in the $\boldsymbol{L}^1$-norm but the accuracy reduces to first order in the $\boldsymbol{L}^\infty$-norm (for smooth solutions). This order barrier can be overcome by slightly relaxing the lower bound defined in Lemma 83.14, i.e., $\Psi_i^{\min} := \min_{j \in \mathcal{I}(i)} \Psi(\overline{\mathsf{U}}_{ij}^n)$. This can be done efficiently while preserving the global invariant domain properties by proceeding as in Guermond et al. [169, §4.7.1], [171, §7.6]. □

Exercises

Exercise 83.1 (Dispersion error). Let $u(x,t)$ be a smooth function satisfying $\partial_t u + \beta \partial_x u = 0$, $x \in D := (0,1)$, $t > 0$, where $\beta \in \mathbb{R}$. Let $I \in \mathbb{N} \backslash \{0\}$ and consider the uniform mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, $\forall i \in \{1{:}I{-}1\}$, with size $h := \frac{1}{I-1} = x_{i+1} - x_i$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$ and let $\{\varphi_i\}_{i \in \mathcal{A}_h}$, $\mathcal{A}_h = \{1{:}I\}$, be the associated global Lagrange shape functions. (i) Compute the coefficients of the consistent mass matrix, $\mathcal{M}$, and the coefficients of the lumped mass matrix, $\overline{\mathcal{M}}$. (ii) Keep the time continuous and write the Galerkin approximation using the lumped mass matrix of the Cauchy problem (with the boundary condition equal to the initial condition as above) for a test function φ_i, $\forall i \in \mathcal{A}_h^\circ = \{2{:}I{-}1\}$. (iii) Let $\mathcal{I}_h^{\mathrm{L}}(u)$ be the Lagrange approximation of u. Using Taylor expansions, estimate (informally) the leading term in the consistency error $R_i^{\mathrm{L}}(t) := \frac{1}{\int_D \varphi_i \,\mathrm{d}x} \overline{\mathcal{M}} \partial_t u(x_i,t) + \int_D (\beta \partial_x \mathcal{I}_h^{\mathrm{L}}(u)) \varphi_i \,\mathrm{d}x$, $\forall i \in \mathcal{A}_h^\circ$. (iv) Keep the time continuous and write the Galerkin approximation using the consistent mass matrix of the Cauchy problem for a test function φ_i, $\forall i \in \mathcal{A}_h^\circ$. (v) Using Taylor expansions, estimate (informally) the leading term in the consistency error $R_i(t) := \frac{1}{\int_D \varphi_i \,\mathrm{d}x} \int_D \big(\partial_t(\mathcal{I}_h^{\mathrm{L}}(u)) + \beta \partial_x(\mathcal{I}_h^{\mathrm{L}}(u))\big)\varphi_i \,\mathrm{d}x$, $\forall i \in \mathcal{A}_h^\circ$. (*Hint*: $u(x_i \pm h, t) = u(x_i) \pm h\partial_x u(x,t) + \frac{1}{2}h^2 \partial_{xx} u(x_i,t) \pm \frac{1}{6}h^3 \partial_{xxx} u(x_i,t) + \frac{1}{24}h^4 \partial_{xxxx} u(x_i,t) \pm \frac{1}{120}h^5 \partial_{xxxxx} u(x_i,t) + \mathcal{O}(h^6)$.)

Exercise 83.2 (FCT counterexample). Consider 1D Burgers' equation, $\boldsymbol{f}(u) := f(u)\boldsymbol{e}_x$, $f(u) := \frac{1}{2}u^2$, $D := (-1,1)$, with initial data $u_0(x) := -1$ if $x \le 0$ and $u_0(x) := 1$ otherwise. Let $I \ge 3$ be an odd number, and consider the (nonuniform) mesh $\mathcal{T}_h$ composed of the cells $[x_i, x_{i+1}]$, where the nodes x_i, $\forall i \in \mathcal{A}_h := \{1{:}I\}$, are such that $-1 =: x_1 < \cdots < x_I := 1$ and $x_{I'} \le 0 < x_{I'+1}$ with $I' := \frac{I+1}{2}$. Let $P_1^{\mathrm{g}}(\mathcal{T}_h)$ be the finite element space composed of continuous piecewise linear functions on $\mathcal{T}_h$ and let $\{\varphi_i\}_{i \in \mathcal{A}_h}$ be the associated global Lagrange shape functions. (i) Compute $\boldsymbol{c}_{i,i-1}$, $\boldsymbol{c}_{i,i}$, $\boldsymbol{c}_{i,i+1}$, and m_i for all $i \in \mathcal{A}_h^\circ := \{2{:}I{-}1\}$. (ii) Let $u_h^0 := \sum_{i \in \mathcal{A}_h} \mathsf{U}_i^0 \varphi_i(x)$ with $\mathsf{U}_i^0 := -1$ if $i \le I'$ and $\mathsf{U}_i^0 := 1$ if $i > I'$. Compute the Galerkin solution at $t := \tau$ using the lumped mass

matrix, say $u_h^{\mathrm{H},1}$. (iii) What is the maximum wave speed in the Riemann problem with the data $(-1,1)$? (iv) Compute the low-order solution at $t := \tau$, say $u_h^{\mathrm{L},1}$. (v) Using the notation of the FCT limiting, compute a_{ij} for all $i \in \mathcal{A}_h^\circ$ and all $j \in \mathcal{I}(i) := \{i-1, i, i+1\}$. (vi) Show that $\ell_{ij} = 1$ for all $i \in \mathcal{A}_h^\circ$ and all $j \in \mathcal{I}(i)$. (vii) Does the approximate solution converge to the entropy solution?

Exercise 83.3 (Quasiconcavity). (i) Let $B \subset \mathbb{R}^m$ be a convex set. Show that a function $\Psi : B \to \mathbb{R}$ is quasiconcave iff for every finite set $\{\mathbf{U}_i\}_{i\in I} \subset B$ and all numbers $\{\theta_i\}_{i\in I} \subset [0,1]$ with $\sum_{i\in I}\theta_i = 1$, one has $\Psi(\sum_{i\in I}\theta_i\mathbf{U}_i) \geq \min_{i\in I}\Psi(\mathbf{U}_i)$. (ii) Let $\mathcal{A} \subset \mathbb{R}^m$ be a convex set. Let $\phi : \mathcal{A} \to \mathbb{R}$ be a quasiconcave function. Let $\boldsymbol{z} \in \mathbb{R}^m$, and let $L : \mathcal{A} \to \mathbb{R}$ be defined by $L(\boldsymbol{u}) := \boldsymbol{z}\cdot\boldsymbol{u}$ for all $\boldsymbol{u} \in \mathcal{A}$. Let $\phi : \mathcal{A} \to \mathbb{R}$ be a continuous function. Let $B := \{\boldsymbol{u} \in \mathcal{A} \mid L(\boldsymbol{u}) > 0\}$ and assume that $B \neq \emptyset$. Assume that $\psi : B \to \mathbb{R}$ defined by $\psi(\boldsymbol{u}) := L(\boldsymbol{u})\phi(\boldsymbol{u})$ is concave. Prove that $\phi_{|B} : B \to \mathbb{R}$ is quasiconcave. (A first example for the Euler equations is $B := \mathcal{A} = \{\boldsymbol{u} \in \mathbb{R}^m \mid \rho > 0\}$ with $L(\boldsymbol{u}) := \rho$, $\phi(\boldsymbol{u}) := e(\boldsymbol{u}) := \rho^{-1}E - \frac{1}{2}\rho^{-2}\boldsymbol{m}^2$, where $e(\boldsymbol{u})$ is the specific internal energy. Another example is $B := \mathcal{A} = \{\boldsymbol{u} \in \mathbb{R}^m \mid \rho > 0, e(\boldsymbol{u}) > 0\}$, $\phi(\boldsymbol{u}) := \Phi(\boldsymbol{u})$, where $\Phi(\boldsymbol{u})$ is the specific entropy.)

Exercise 83.4 (Harten's lemma). (i) Consider the following scheme for scalar conservation equations $\mathsf{U}_i^{n+1} = \mathsf{U}_i^n - C_{i-1}^n(\mathsf{U}_i^n - \mathsf{U}_{i-1}^n) + D_i^n(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n)$ for all $i \in \mathbb{Z}$. Assume that $0 \leq C_i^n$, $0 \leq D_i^n$, $C_i^n + D_i^n \leq 1$ for all $i \in \mathbb{Z}$. Let $|\mathsf{V}|_{\mathrm{TV}} := \sum_{i\in\mathbb{Z}}|\mathsf{V}_{i+1} - \mathsf{V}_i|$ be the total variation of $\mathsf{V} \in \mathbb{R}^{\mathbb{Z}}$. Prove that the above algorithm is *total variation diminishing* (TVD), i.e., $|\mathsf{U}^{n+1}|_{\mathrm{TV}} \leq |\mathsf{U}^n|_{\mathrm{TV}}$. (ii) Consider the method described in (81.9)-(81.10) in dimension one. Assume that $\mathcal{I}(i) = \{i-1, i, i+1\}$ and that the mesh is infinite in both directions. Show that the method can be put into the above form and satisfies the above assumptions if $4\tau\sup_{i\in\mathbb{Z}}\frac{|d_{ii}^n|}{m_i} \leq 1$. (*Hint*: see Exercise 79.4.)

Exercise 83.5 (Lax–Wendroff). Let u be a smooth solution to the scalar transport equation $\partial_t u + a\partial_x u = 0$ with $a \in \mathbb{R}_+$. (i) Using finite Taylor expansions, show that $u(x, t_{n+1}) = u(x, t_n) - \tau a\partial_x u(x, t_n) + \frac{a^2\tau^2}{2}\partial_{xx}u(x, t_n) + \mathcal{O}(\tau^3)$. (ii) Consider now the time-stepping algorithm consisting of setting $u^0 := u_0$ and for all $n \geq 0$, $u^{n+1}(x) := u^n(x) - \tau a\partial_x u^n(x) + \frac{a^2\tau^2}{2}\partial_{xx}u^n(x)$. What is the (informal) order of accuracy of this method with respect to τ? (iii) Let $\mathcal{T}_h$ be a uniform mesh in $D := (0,1)$ with grid points $x_i := (i-1)h$, $\forall i \in \mathcal{A}_h := \{1{:}I\}$, $h := \frac{1}{I-1}$. Let $\{\varphi_i\}_{i\in\mathcal{A}_h}$ be the piecewise linear Lagrange shape functions associated with the grid points $\{x_i\}_{i\in\mathcal{A}_h}$. Let x_i be an interior node, i.e., $i \in \mathcal{A}_h^\circ := \{2{:}I{-}1\}$. Write the equation corresponding to the Galerkin approximation using the lumped mass matrix of the equation $u^{n+1}(x) = u^n(x) - \tau a\partial_x u^n(x) + \frac{a^2\tau^2}{2}\partial_{xx}u^n(x)$ with homogeneous Neumann boundary conditions using the test function φ_i, where both u^{n+1} and u^n are approximated in $P_1^{\mathrm{g}}(\mathcal{T}_h) := \mathrm{span}\{\varphi_i\}_{i\in\mathcal{A}_h}$. (iv) What is the (informal) order of accuracy of this method with respect to τ and h? (v) Let $u_h^{\mathrm{L},n+1} := \sum_{i\in\mathcal{A}_h}\mathsf{U}_i^{\mathrm{L},n+1}\varphi_i$ be the first-order approximation of u using (81.9)-(81.10). Show that $m_i\mathsf{U}_i^{n+1} = m_i\mathsf{U}_i^{\mathrm{L},n+1} + \frac{a\tau}{2}(\lambda - 1)(\mathsf{U}_{i+1}^n - \mathsf{U}_i^n) + \frac{a\tau}{2}(\lambda - 1)(\mathsf{U}_{i-1}^n - \mathsf{U}_i^n)$, where $\gamma := \frac{a\tau}{h}$. *Note*: the scheme is now ready for FCT limiting. Actually, there exists in the literature a plethora of limiting techniques (like FCT) that, after applying the limiter, make the scheme TVD in the sense of Exercise 83.4; see Sweby [267].

Bibliography

[1] R. Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *J. Sci. Comput.*, 73(2-3):461–494, 2017. pages 325

[2] N. Ahmed and G. Matthies. Higher order continuous Galerkin-Petrov time stepping schemes for transient convection-diffusion-reaction equations. *ESAIM Math. Model. Numer. Anal.*, 49(5):1429–1450, 2015. pages 167

[3] N. Ahmed, S. Becher, and G. Matthies. Higher-order discontinuous Galerkin time stepping and local projection stabilization techniques for the transient Stokes problem. *Comput. Methods Appl. Mech. Engrg.*, 313:28–52, 2017. pages 209

[4] N. Ahmed, T. Chacón Rebollo, V. John, and S. Rubino. Analysis of a full space-time discretization of the Navier-Stokes equations by a local projection stabilization method. *IMA J. Numer. Anal.*, 37(3):1437–1467, 2017. pages 209

[5] M. Ainsworth. Pyramid algorithms for Bernstein-Bézier finite elements of high, nonuniform order in any dimension. *SIAM J. Sci. Comput.*, 36(2):A543–A569, 2014. pages 296, 299

[6] M. Ainsworth. Dispersive behaviour of high order finite element schemes for the one-way wave equation. *J. Comput. Phys.*, 259:1–10, 2014. pages 324

[7] G. Akrivis and C. Makridakis. Galerkin time-stepping methods for nonlinear parabolic equations. *M2AN Math. Model. Numer. Anal.*, 38(2):261–289, 2004. pages 153

[8] H. Amann. Compact embeddings of vector-valued Sobolev and Besov spaces. *Glas. Mat. Ser. III*, 35(55)(1):161–177, 2000. pages 101

[9] R. Andreev. *Stability of space-time Petrov–Galerkin discretizations for parabolic evolution equations*. PhD thesis, ETH Zürich, 2012. pages 162

[10] R. Andreev. Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal.*, 33(1):242–260, 2013. pages 162, 183

[11] N. Antonić and K. Burazin. Graph spaces of first-order linear partial differential operators. *Math. Commun.*, 14(1):135–155, 2009. pages 9

[12] N. Antonić and K. Burazin. Intrinsic boundary conditions for Friedrichs systems. *Comm. Partial Differential Equations*, 35(9):1690–1715, 2010. pages 9

[13] N. Antonić and K. Burazin. Boundary operator from matrix field formulation of boundary conditions for Friedrichs systems. *J. Differential Equations*, 250(9):3630–3651, 2011. pages 9

[14] D. Arndt, H. Dallmann, and G. Lube. Local projection FEM stabilization for the time-dependent incompressible Navier-Stokes problem. *Numer. Methods Partial Differential Equations*, 31(4):1224–1250, 2015. pages 209

[15] J.-P. Aubin. Un théorème de compacité. *C. R. Acad. Sci. Paris, Sér. I*, 256:5042–5044, 1963. pages 101

[16] J.-P. Aubin. *Applied functional analysis*. Pure and Applied Mathematics. Wiley-Interscience, New York, NY, second edition, 2000. With exercises by B. Cornet and J.-M. Lasry, Translated from the French by C. Labrousse. pages 7, 8

[17] B. Ayuso and L. D. Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 47(2):1391–1420, 2009. pages 70

[18] P. Azerad. *Analyse des équations de Navier–Stokes en bassin peu profond et de l'équation de transport*. Ph. D. thesis, Univeristé de Neuchâtel, Switzerland, 1995. pages 70

[19] A. K. Aziz and P. Monk. Continuous finite elements in space and time for the heat equation. *Math. Comp.*, 52(186):255–274, 1989. pages 167, 174

[20] A. K. Aziz, R. B. Kellogg, and A. B. Stephens. Least-squares methods for elliptic systems. *Math. Comp.*, 44(169):53–70, 1985. pages 17

[21] G. A. Baker, J. H. Bramble, and V. Thomée. Single step Galerkin approximations for parabolic problems. *Math. Comp.*, 31(140):818–847, 1977. pages 145

[22] J. M. Ball. Strongly continuous semigroups, weak solutions, and the variation of constants formula. *Proc. Amer. Math. Soc.*, 63(2):370–373, 1977. pages 238

[23] C. Bardos, A. Y. le Roux, and J.-C. Nédélec. First order quasilinear equations with boundary conditions. *Comm. Partial Differential Equations*, 4(9):1017–1034, 1979. pages 269

[24] G. R. Barrenechea and F. Valentin. Consistent local projection stabilized finite element methods. *SIAM J. Numer. Anal.*, 48(5):1801–1825, 2010. pages 85

[25] G. R. Barrenechea, E. Burman, and F. Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.*, 135(2):521–545, 2017. pages 316

[26] M. Bause, F. A. Radu, and U. Köcher. Error analysis for discretizations of parabolic problems using continuous finite elements in time and mixed finite elements in space. *Numer. Math.*, 137(4):773–818, 2017. pages 167

[27] Y. Bazilevs and T. J. R. Hughes. Weak imposition of Dirichlet boundary conditions in fluid mechanics. *Comput. & Fluids*, 36(1):12–26, 2007. pages 62

[28] R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, 38(4):173–199, 2001. pages 41, 82

[29] H. Beirão da Veiga. On a stationary transport equation. *Ann. Univ. Ferrara Sez. VII*, 32: 79–91, 1986. pages 70

[30] A. Bensalah, P. Joly, and J.-F. Mercier. Well-posedness of a generalized time-harmonic transport equation for acoustics in flow. *Math. Methods Appl. Sci.*, 41(8):3117–3137, 2018. pages 70

[31] C. Berthon, F. Coquel, and P. G. LeFloch. Why many theories of shock waves are necessary: kinetic relations for non-conservative systems. *Proc. Roy. Soc. Edinburgh Sect. A*, 142(1): 1–37, 2012. pages 283

[32] S. Bianchini and A. Bressan. Vanishing viscosity solutions of nonlinear hyperbolic systems. *Ann. of Math. (2)*, 161(1):223–342, 2005. pages 284, 290

[33] P. B. Bochev. Experiences with negative norm least-square methods for the Navier-Stokes equations. *Electron. Trans. Numer. Anal.*, 6:44–62, 1997. pages 17

[34] P. B. Bochev. Negative norm least-squares methods for the velocity-vorticity-pressure Navier-Stokes equations. *Numer. Methods Partial Differential Equations*, 15(2):237–256, 1999. pages 17

[35] P. B. Bochev, C. R. Dohrmann, and M. D. Gunzburger. Stabilization of low-order mixed finite elements for the Stokes equations. *SIAM J. Numer. Anal.*, 44(1):82–101, 2006. pages 82

[36] S. Bochner and A. E. Taylor. Linear functionals on certain spaces of abstractly-valued functions. *Ann. of Math. (2)*, 39(4):913–944, 1938. pages 96

[37] T. Boiveau, V. Ehrlacher, A. Ern, and A. Nouy. Low-rank approximation of linear parabolic equations by space-time tensor Galerkin methods. *ESAIM Math. Model. Numer. Anal.*, 53 (2):635–658, 2019. pages 162

[38] A. Bonito, J.-L. Guermond, and B. Popov. Stability analysis of explicit entropy viscosity methods for non-linear scalar conservation equations. *Math. Comp.*, 83(287):1039–1062, 2014. pages 261

[39] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. **11** (1973), no. 1, 38–69]. *J. Comput. Phys.*, 135(2):170–186, 1997. pages 327

[40] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, Switzerland, 2004. pages 281

[41] M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, 43(6):2544–2566, 2006. pages 41, 62

[42] M. Braack, E. Burman, V. John, and G. Lube. Stabilized finite element methods for the generalized Oseen problem. *Comput. Methods Appl. Mech. Engrg.*, 196(4-6):853–866, 2007. pages 73, 75, 82

[43] J. H. Bramble and J. E. Pasciak. Least-squares methods for Stokes equations based on a discrete minus one inner product. *J. Comput. Appl. Math.*, 74(1-2):155–173, 1996. pages 17

[44] J. H. Bramble and A. H. Schatz. Rayleigh-Ritz-Galerkin-methods for Dirichlet's problem using subspaces without boundary conditions. *Comm. Pure Appl. Math.*, 23:653–675, 1970. pages 17

[45] J. H. Bramble and A. H. Schatz. Least squares for $2m$th order elliptic boundary-value problems. *Math. Comp.*, 25:1–32, 1971. pages 17

[46] J. H. Bramble and T. Sun. A negative-norm least squares method for Reissner-Mindlin plates. *Math. Comp.*, 67(223):901–916, 1998. pages 17

[47] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.*, 66(219):935–955, 1997. pages 17

[48] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak. Least-squares methods for linear elasticity based on a discrete minus one inner product. *Comput. Methods Appl. Mech. Engrg.*, 191 (8-10):727–744, 2001. pages 17

[49] S. C. Brenner. Korn's inequalities for piecewise H^1 vector fields. *Math. Comp.*, 73(247): 1067–1087, 2004. pages 83, 88

[50] A. Bressan. *Hyperbolic systems of conservation laws: The one-dimensional Cauchy problem*, volume 20 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2000. pages 290, 304

[51] A. Bressan. Hyperbolic conservation laws: an illustrated tutorial. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 157–245. Springer, Heidelberg, 2013. pages 281

[52] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, NY, 2011. pages 101, 110, 115, 236, 238

[53] F. Brezzi and J. Pitkäranta. On the stabilization of finite element approximations of the Stokes equations. In *Efficient solutions of elliptic systems (Kiel, 1984)*, volume 10 of *Notes Numer. Fluid Mech.*, pages 11–19. Friedr. Vieweg, Braunschweig, 1984. pages 75

[54] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin methods for first-order hyperbolic problems. *Math. Models Methods Appl. Sci.*, 14(12):1893–1903, 2004. pages 57

[55] A. N. Brooks and T. J. R. Hughes. Streamline Upwind/Petrov–Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32:199–259, 1982. pages 20

[56] D. L. Brown, R. Cortez, and M. L. Minion. Accurate projection methods for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 168(2):464–499, 2001. pages 218

[57] E. Burman. A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, 43(5):2012–2033, 2005. pages 32, 62

[58] E. Burman. On nonlinear artificial viscosity, discrete maximum principle and hyperbolic conservation laws. *BIT*, 47(4):715–733, 2007. pages 316

[59] E. Burman. Consistent SUPG-method for transient transport problems: stability and convergence. *Comput. Methods Appl. Mech. Engrg.*, 199(17-20):1114–1123, 2010. pages 241

[60] E. Burman. Robust error estimates in weak norms for advection dominated transport problems with rough data. *Math. Models Methods Appl. Sci.*, 24(13):2663–2684, 2014. pages 70

[61] E. Burman and A. Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.*, 74(252):1637–1652, 2005. pages 316

[62] E. Burman and A. Ern. Continuous interior penalty *hp*-finite element methods for advection and advection-diffusion equations. *Math. Comp.*, 76(259):1119–1140, 2007. pages 32

[63] E. Burman and A. Ern. A continuous finite element method with face penalty to approximate Friedrichs' systems. *M2AN Math. Model. Numer. Anal.*, 41(1):55–76, 2007. pages 32

[64] E. Burman and A. Ern. Implicit-explicit Runge-Kutta schemes and finite elements with symmetric stabilization for advection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(4):681–707, 2012. pages 253

[65] E. Burman and M. A. Fernández. Continuous interior penalty finite element method for the time-dependent Navier-Stokes equations: space discretization and convergence. *Numer. Math.*, 107(1):39–77, 2007. pages 209

[66] E. Burman and M. A. Fernández. Finite element methods with symmetric stabilization for the transient convection-diffusion-reaction equation. *Comput. Methods Appl. Mech. Engrg.*, 198(33-36):2508–2519, 2009. pages 253

[67] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004. pages 32, 62

[68] E. Burman and P. Hansbo. Edge stabilization for the generalized Stokes problem: a continuous interior penalty method. *Comput. Methods Appl. Mech. Engrg.*, 195(19-22):2393–2410, 2006. pages 82

[69] E. Burman and F. Schieweck. Local CIP stabilization for composite finite elements. *SIAM J. Numer. Anal.*, 54(3):1967–1992, 2016. pages 35, 81

[70] E. Burman and B. Stamm. Minimal stabilization for discontinuous Galerkin finite element methods for hyperbolic problems. *J. Sci. Comput.*, 33:183–208, 2007. pages 57

[71] E. Burman and B. Stamm. Bubble stabilized discontinuous Galerkin method for Stokes' problem. *Math. Models Methods Appl. Sci.*, 20(2):297–313, 2010. pages 88

[72] E. Burman, M. A. Fernández, and P. Hansbo. Continuous interior penalty finite element method for Oseen's equations. *SIAM J. Numer. Anal.*, 44(3):1248–1274, 2006. pages 62, 82

[73] E. Burman, J. Guzmán, and D. Leykekhman. Weighted error estimates of the continuous interior penalty method for singularly perturbed problems. *IMA J. Numer. Anal.*, 29(2): 284–314, 2009. pages 68

[74] E. Burman, A. Ern, and M. A. Fernández. Explicit Runge-Kutta schemes and finite elements with symmetric stabilization for first-order linear PDE systems. *SIAM J. Numer. Anal.*, 48 (6):2019–2042, 2010. pages 265, 266

[75] E. Burman, A. Quarteroni, and B. Stamm. Interior penalty continuous and discontinuous finite element approximations of hyperbolic equations. *J. Sci. Comput.*, 43(3):293–312, 2010. pages 57

[76] E. Burman, A. Ern, and M. A. Fernández. Fractional-step methods and finite elements with symmetric stabilization for the transient Oseen problem. *ESAIM Math. Model. Numer. Anal.*, 51(2):487–507, 2017. pages 221

[77] J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18:50–64, 1964. pages 168, 176, 257

[78] J. C. Butcher. A history of Runge-Kutta methods. *Appl. Numer. Math.*, 20:247–260, 1996. Selected keynote papers presented at 14th IMACS World Congress (Atlanta, GA, 1994). pages 168

[79] P. Cantin. Well-posedness of the scalar and the vector advection-reaction problems in Banach graph spaces. *C. R. Math. Acad. Sci. Paris*, 355(8):892–902, 2017. pages 70

[80] P. Cantin and A. Ern. An edge-based scheme on polyhedral meshes for vector advection-reaction equations. *ESAIM Math. Model. Numer. Anal.*, 51(5):1561–1581, 2017. pages 70

[81] E. Chiodaroli, C. De Lellis, and O. Kreml. Global ill-posedness of the isentropic system of gas dynamics. *Comm. Pure Appl. Math.*, 68(7):1157–1190, 2015. pages 284

[82] A. J. Chorin. A numerical method for solving incompressible viscous flow problems. *J. Comp. Phys.*, 2:12–26, 1967. pages 227

[83] A. J. Chorin. Numerical solution of the Navier–Stokes equations. *Math. Comp.*, 22:745–762, 1968. pages 213, 216

[84] A. J. Chorin. On the convergence of discrete approximations to the Navier–Stokes equations. *Math. Comp.*, 23:341–353, 1969. pages 213, 216

[85] M. A. Christon, M. J. Martinez, and T. E. Voth. Generalized Fourier analyses of the advection-diffusion equation-Part I: One-dimensional domains. *International Journal for Numerical Methods in Fluids*, 45(8):839–887, 2004. pages 324

[86] K. Chrysafinos and L. S. Hou. Error estimates for semidiscrete finite element approximations of linear and semilinear parabolic equations under minimal regularity assumptions. *SIAM J. Numer. Anal.*, 40(1):282–306, 2002. pages 185

[87] K. Chrysafinos and N. J. Walkington. Error estimates for the discontinuous Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 44(1):349–366, 2006. pages 153

[88] K. N. Chueh, C. C. Conley, and J. A. Smoller. Positively invariant regions for systems of nonlinear diffusion equations. *Indiana Univ. Math. J.*, 26(2):373–392, 1977. pages 291

[89] B. Cockburn, G. Kanschat, D. Schötzau, and C. Schwab. Local discontinuous Galerkin methods for the Stokes system. *SIAM J. Numer. Anal.*, 40(1):319–343, 2002. pages 88

[90] R. Codina. Stabilized finite element approximation of transient incompressible flows using orthogonal subscales. *Comput. Methods Appl. Mech. Engrg.*, 191(39-40):4295–4321, 2002. pages 41, 82, 209

[91] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100(1):32–74, 1928. pages 255, 300

[92] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM J. Res. Develop.*, 11:215–234, 1967. pages 255, 300

[93] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Cambridge Philos. Soc.*, 43:50–67, 1947. pages 142

[94] M. Crouzeix. Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques. *Numer. Math.*, 35(3):257–276, 1980. pages 138

[95] M. Crouzeix and P.-A. Raviart. Approximation des équations d'évolution linéaires par des méthodes à pas multiples. *C. R. Acad. Sci. Paris Sér. A-B*, 28(6):Aiv, A367–A370, 1976. pages 138

[96] C. M. Dafermos. Polygonal approximations of solutions of the initial value problem for a conservation law. *J. Math. Anal. Appl.*, 38:33–41, 1972. pages 276

[97] C. M. Dafermos. *Hyperbolic conservation laws in continuum physics*, volume 325 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2000. pages 290

[98] G. Dal Maso, P. G. Lefloch, and F. Murat. Definition and weak stability of nonconservative products. *J. Math. Pures Appl. (9)*, 74(6):483–548, 1995. pages 283

[99] H. Dallmann, D. Arndt, and G. Lube. Local projection stabilization for the Oseen problem. *IMA J. Numer. Anal.*, 36(2):796–823, 2016. pages 209

[100] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 5. Evolution problems, I.* Springer-Verlag, Berlin, Germany, 1992. pages 99, 101, 108, 181

[101] C. De Lellis and L. Székelyhidi, Jr. On admissibility criteria for weak solutions of the Euler equations. *Arch. Ration. Mech. Anal.*, 195(1):225–260, 2010. pages 284

[102] M. Delfour, W. Hager, and F. Trochu. Discontinuous Galerkin methods for ordinary differential equations. *Math. Comp.*, 36(154):455–473, 1981. pages 153

[103] P. Deuring, R. Eymard, and M. Mildner. L^2-stability independent of diffusion for a finite element-finite volume discretization of a linear convection-diffusion equation. *SIAM J. Numer. Anal.*, 53(1):508–526, 2015. pages 70

[104] A. Devinatz, R. Ellis, and A. Friedman. The asymptotic behavior of the first real eigenvalue of second order elliptic operators with a small parameter in the highest derivatives. II. *Indiana Univ. Math. J.*, 23:991–1011, 1973–1974. pages 68, 70

[105] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications [Mathematics & Applications]*. Springer-Verlag, Berlin, 2012. pages 62, 88

[106] D. A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection. *SIAM J. Numer. Anal.*, 46(2):805–831, 2008. pages 62

[107] J. Diestel. *Sequences and series in Banach spaces*, volume 92 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1984. pages 94, 95

[108] C. R. Dohrmann and P. B. Bochev. A stabilized finite element method for the Stokes problem based on polynomial pressure projections. *Internat. J. Numer. Methods Fluids*, 46(2):183–201, 2004. pages 82

[109] J. Douglas, Jr. and T. Dupont. Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 7:575–626, 1970. pages 145

[110] J. Douglas, Jr., T. Dupont, and M. F. Wheeler. A quasi-projection analysis of Galerkin methods for parabolic and hyperbolic equations. *Math. Comp.*, 32(142):345–362, 1978. pages 145

[111] A. V. C. Duarte, E. G. D. do Carmo, and F. A. Rochinha. Consistent discontinuous finite elements in elastodynamics. *Comput. Methods Appl. Mech. Engrg.*, 190(1-2):193–223, 2000. pages 83, 88

[112] F. Dubois and P. LeFloch. Boundary conditions for nonlinear hyperbolic systems of conservation laws. *J. Differential Equations*, 71(1):93–122, 1988. pages 281

[113] T. Dupont. Mesh modification for evolution equations. *Math. Comp.*, 39(159):85–107, 1982. pages 125, 185

[114] A. V. Džiškariani. The least square and Bubnov-Galerkin methods. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 8:1110–1116, 1968. pages 17

[115] W. E and J. G. Liu. Projection method I: Convergence and numerical boundary layers. *SIAM J. Numer. Anal.*, 32:1017–1057, 1995. pages 218

[116] K. Eriksson, C. Johnson, and V. Thomée. Time discretization of parabolic problems by the discontinuous Galerkin method. *RAIRO Modél. Math. Anal. Numér.*, 19(4):611–643, 1985. pages 153

[117] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004. pages 181

[118] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006. pages 7, 11, 57

[119] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 44(6):2363–2388, 2006. pages 7, 57

[120] A. Ern and J.-L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. III. Multi-field theories with partial coercivity. *SIAM J. Numer. Anal.*, 46(2):776–804, 2008. pages 7, 57

[121] A. Ern and J.-L. Guermond. Weighting the edge stabilization. *SIAM J. Numer. Anal.*, 51 (3):1655–1677, 2013. pages 32

[122] A. Ern and F. Schieweck. Discontinuous Galerkin method in time combined with a stabilized finite element method in space for linear first-order PDEs. *Math. Comp.*, 85(301):2099–2129, 2016. pages 154, 253

[123] A. Ern, J.-L. Guermond, and G. Caplain. An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems. *Comm. Partial Differ. Eq.*, 32:317–341, 2007. pages 9

[124] A. Ern, A. F. Stephansen, and P. Zunino. A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity. *IMA J. Numer. Anal.*, 29(2):235–256, 2009. pages 62

[125] A. Ern, I. Smears, and M. Vohralík. Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems. *SIAM J. Numer. Anal.*, 55(6):2811–2834, 2017. pages 154, 180, 181

[126] L. Ferracina and M. N. Spijker. An extension and analysis of the Shu-Osher representation of Runge-Kutta methods. *Math. Comp.*, 74(249):201–219, 2005. pages 307, 311, 312

[127] M. S. Floater. Generalized barycentric coordinates and applications. *Acta Numer.*, 24: 161–214, 2015. pages 317

[128] M. S. Floater, K. Hormann, and G. Kós. A general construction of barycentric coordinates over convex polygons. *Advances in Computational Mathematics*, 24(1):311–331, 2006. pages 317

[129] L. P. Franca and S. L. Frey. Stabilized finite element methods: II. The incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 99:209–233, 1992. pages 75

[130] H. Frid. Maps of convex sets and invariant regions for finite-difference systems of conservation laws. *Arch. Ration. Mech. Anal.*, 160(3):245–269, 2001. pages 291

[131] K. O. Friedrichs. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11:333–418, 1958. pages 1

[132] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proc. Nat. Acad. Sci. U.S.A.*, 68:1686–1688, 1971. pages 286

[133] B. García-Archilla and E. S. Titi. Postprocessing the Galerkin method: the finite-element case. *SIAM J. Numer. Anal.*, 37(2):470–499, 2000. pages 117

[134] B. García-Archilla, J. Novo, and E. S. Titi. Postprocessing the Galerkin method: a novel approach to approximate inertial manifolds. *SIAM J. Numer. Anal.*, 35(3):941–972, 1998. pages 117

[135] V. Girault, B. Rivière, and M. F. Wheeler. A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier-Stokes problems. *Math. Comp.*, 74 (249):53–84, 2005. pages 88

[136] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.*, 18:697–715, 1965. pages 284

[137] K. Goda. A multistep technique with implicit difference schemes for calculating two- or three-dimensional cavity flows. *J. Comput. Phys.*, 30:76–95, 1979. pages 216

[138] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996. pages 281, 286, 289, 290, 303

[139] S. K. Godunov. Themodynamics of gases and differential equations. *Uspehi Mat. Nauk*, 14 (5 (89)):97–116, 1959. pages 286

[140] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001. pages 309

[141] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu. High order strong stability preserving time discretizations. *J. Sci. Comput.*, 38(3):251–289, 2009. pages 307, 310

[142] J.-L. Guermond. Remarques sur les méthodes de projection pour l'approximation des équations de Navier-Stokes. *Numer. Math.*, 67(4):465–473, 1994. pages 196, 221

[143] J.-L. Guermond. Some practical implementations of projection methods for Navier-Stokes equations. *RAIRO Modél. Math. Anal. Numér.*, 30:637–667, 1996. Also in *C. R. Acad. Sci. Paris, Série I*, 319:887–892, 1994. pages 220, 221

[144] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33(6):1293–1316, 1999. pages 40, 41, 43, 62

[145] J.-L. Guermond. Un résultat de convergence d'ordre deux en temps pour l'approximation des équations de Navier-Stokes par une technique de projection incrémentale. *M2AN Math. Model. Numer. Anal.*, 33(1):169–189, 1999. Also in *C. R. Acad. Sci. Paris, Série I*, 325:1329–1332, 1997. pages 218, 220

[146] J.-L. Guermond. Stabilisation par viscosité de sous-maille pour l'approximation de Galerkin des opérateurs linéaires monotones. *C. R. Acad. Sci. Paris Sér. I Math.*, 328(7):617–622, 1999. pages 40, 41

[147] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of monotone operators. *ZAMM-Z. Angew. Math. Mech.*, 79(S1):29–32, 1999. pages 41

[148] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear monotone operators. *IMA J. Numer. Anal.*, 21:165–197, 2001. pages 40, 41

[149] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear contraction semigroups of class C^0 in Hilbert spaces. *Numer. Methods Partial Differential Equations*, 17(1): 1–25, 2001. pages 41, 253

[150] J.-L. Guermond and P. Minev. High-order time stepping for the incompressible Navier-Stokes equations. *SIAM J. Sci. Comput.*, 37(6):A2656–A2681, 2015. pages 223

[151] J.-L. Guermond and P. Minev. High-order time stepping for the Navier-Stokes equations with minimal computational complexity. *J. Comput. Appl. Math.*, 310:92–103, 2017. pages 223

[152] J.-L. Guermond and P. Minev. High-order adaptive time stepping for the incompressible Navier-Stokes equations. *SIAM J. Sci. Comput.*, 41(2):A770–A788, 2019. pages 223, 229, 230

[153] J.-L. Guermond and M. Nazarov. A maximum-principle preserving C^0 finite element method for scalar conservation equations. *Comput. Methods Appl. Mech. Engrg.*, 272:198–213, 2013. pages 295

[154] J.-L. Guermond and R. Pasquetti. A correction technique for the dispersive effects of mass lumping for transport problems. *Comput. Methods Appl. Mech. Engrg.*, 253:186–198, 2013. pages 324, 325

[155] J.-L. Guermond and B. Popov. Viscous regularization of the Euler equations and entropy principles. *SIAM J. Appl. Math.*, 74(2):284–305, 2014. pages 286

[156] J.-L. Guermond and B. Popov. Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations. *J. Comput. Phys.*, 321:908–926, 2016. pages 303

[157] J.-L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM Journal on Numerical Analysis*, 54(4):2466–2489, 2016. pages 295, 303, 323, 329

[158] J.-L. Guermond and B. Popov. Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.*, 55(6):3120–3146, 2017. pages 313, 314, 316, 321, 323, 329

[159] J.-L. Guermond and S. Prudhomme. A fully discrete nonlinear Galerkin method for the 3D Navier-Stokes equations. *Numer. Methods Partial Differential Equations*, 24(3):759–775, 2008. pages 117

[160] J.-L. Guermond and L. Quartapelle. On the approximation of the unsteady Navier–Stokes equations by finite element projection methods. *Numer. Math.*, 80(5):207–238, 1998. pages 217, 220, 221

[161] J.-L. Guermond and A. Salgado. A note on the Stokes operator and its powers. *J. Appl. Math. Comput.*, 36(1-2):241–250, 2011. pages 196, 221

[162] J.-L. Guermond and J. Shen. Quelques résultats nouveaux sur les méthodes de projection. *C. R. Acad. Sci. Paris, Sér. I*, 333:1111–1116, 2001. pages 219

[163] J. L. Guermond and J. Shen. On the error estimates for the rotational pressure-correction projection methods. *Math. Comp.*, 73(248):1719–1737, 2004. pages 219

[164] J.-L. Guermond, A. Marra, and L. Quartapelle. Subgrid stabilized projection method for 2D unsteady flows at high Reynolds numbers. *Comput. Methods Appl. Mech. Engrg.*, 195 (44-47):5857–5876, 2006. pages 209

[165] J. L. Guermond, P. Minev, and J. Shen. An overview of projection methods for incompressible flows. *Comput. Methods Appl. Mech. Engrg.*, 195(44-47):6011–6045, 2006. pages 213

[166] J.-L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *J. Comput. Phys.*, 230(11):4248–4267, 2011. pages 323

[167] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang. A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM J. Numer. Anal.*, 52(4):2163–2182, 2014. pages 323

[168] J.-L. Guermond, B. Popov, and Y. Yang. The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations. *Journal of Scientific Computing*, pages 1–9, 2016. pages 297

[169] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas. Second-order invariant domain preserving approximation of the Euler equations using convex limiting. *SIAM J. Sci. Comput.*, 40(5):A3211–A3239, 2018. pages 316, 321, 322, 329, 330, 333

[170] J.-L. Guermond, M. Quezada de Luna, B. Popov, C. E. Kees, and M. W. Farthing. Well-balanced second-order finite element approximation of the shallow water equations with friction. *SIAM J. Sci. Comput.*, 40(6):A3873–A3901, 2018. pages 303, 316

[171] J.-L. Guermond, B. Popov, and I. Tomas. Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Comput. Methods Appl. Mech. Engrg.*, 347:143–175, 2019. pages 298, 321, 329, 332, 333

[172] A. Guillou and J. L. Soulé. La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation. *Rev. Française Informat. Recherche Opérationnelle*, 3(Sér. R-3):17–44, 1969. pages 170

[173] J. Guzmán. Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems. *J. Numer. Math.*, 14(1):41–56, 2006. pages 68

[174] E. Hairer and G. Wanner. Stiff differential equations solved by Radau methods. *J. Comput. Appl. Math.*, 111(1-2):93–111, 1999. pages 155

[175] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations. II. Stiff and Differential-algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Second revised edition, paperback. pages 155, 168, 176, 257

[176] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I. Non-stiff problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. pages 155, 168, 176, 257

[177] P. Hansbo and M. G. Larson. Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method. *Comput. Methods Appl. Mech. Engrg.*, 191 (17-18):1895–1908, 2002. pages 88

[178] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes. I. *SIAM J. Numer. Anal.*, 24(2):279–309, 1987. pages 316

[179] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25(1):35–61, 1983. pages 300

[180] A. Harten, P. D. Lax, C. D. Levermore, and W. J. Morokoff. Convex entropies and hyperbolicity for general Euler equations. *SIAM J. Numer. Anal.*, 35(6):2117–2127, 1998. pages 286, 293

[181] I. Higueras. Representations of Runge-Kutta methods and strong stability preserving methods. *SIAM J. Numer. Anal.*, 43(3):924–948, 2005. pages 307, 311, 312

[182] D. Hoff. A finite difference scheme for a system of two conservation laws with artificial viscosity. *Math. Comp.*, 33(148):1171–1193, 1979. pages 300

[183] D. Hoff. Invariant regions for systems of conservation laws. *Trans. Amer. Math. Soc.*, 289 (2):591–610, 1985. pages 291, 292

[184] H. Holden and N. H. Risebro. *Front tracking for hyperbolic conservation laws*, volume 152 of *Applied Mathematical Sciences*. Springer, Heidelberg, second edition, 2015. pages 275, 276, 281, 289, 290

[185] B. Holm and T. P. Wihler. Continuous and discontinuous Galerkin time stepping methods for nonlinear initial value problems with application to finite time blow-up. *Numer. Math.*, 138(3):767–799, 2018. pages 154, 163, 167

[186] P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002. pages 62

[187] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior penalty method for the indefinite time-harmonic Maxwell equations. *Numer. Math.*, 100:485–518, 2005. pages 58

[188] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Mixed discontinuous Galerkin approximation of the Maxwell operator: the indefinite case. *M2AN Math. Model. Numer. Anal.*, 39(4):727–753, 2005. pages 58

[189] T. J. R. Hughes and L. P. Franca. A new finite element formulation for computational fluid dynamics. VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces. *Comput. Methods Appl. Mech. Engrg.*, 65(1):85–96, 1987. pages 75

[190] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/Least-Squares method for advection-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 73:173–189, 1989. pages 20

[191] B. L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comp.*, 26:415–426, 1972. pages 167

[192] B. L. Hulme. Discrete Galerkin and related one-step methods for ordinary differential equations. *Math. Comp.*, 26:881–891, 1972. pages 153

[193] S. Hussain, F. Schieweck, and S. Turek. Higher order Galerkin time discretizations and fast multigrid solvers for the heat equation. *J. Numer. Math.*, 19(1):41–61, 2011. pages 167, 175

[194] A. Jameson. Origins and further development of the Jameson-Schmidt-Turkel scheme. *AIAA Journal*, 55(5), 2017. pages 316

[195] A. Jameson, W. Schmidt, and E. Turkel. Numerical solution of the Euler equations by finite volume methods using Runge–Kutta time-stepping schemes. In *14th AIAA Fluid and Plasma Dynamics Conference, Palo Alto, CA*, June 1981. AIAA Paper 1981–1259. pages 316

[196] P. Jamet. Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain. *SIAM J. Numer. Anal.*, 15:912–928, 1978. pages 153

[197] M. Jensen. *Discontinuous Galerkin Methods for Friedrichs Systems with Irregular Solutions*. PhD thesis, University of Oxford, UK, `www.comlab.ox.ac.uk/research/na/thesis/thesisjensen.pdf`, 2004. pages 8, 57

[198] B.-N. Jiang. *The least-squares finite element method*. Scientific Computation. Springer-Verlag, Berlin, Germany, 1998. pages 17

[199] V. John. *Finite element methods for incompressible flow problems*, volume 51 of *Springer Series in Computational Mathematics*. Springer, Cham, Switzerland, 2016. pages 209

[200] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173):1–26, 1986. pages 57

[201] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45(1-3):285–312, 1984. pages 20, 68, 153, 241, 253

[202] C. Johnson, A. H. Schatz, and L. B. Wahlbin. Crosswind smear and pointwise errors in streamline diffusion finite element methods. *Math. Comp.*, 49(179):25–38, 1987. pages 68

[203] P. Joly. Some trace theorems in anisotropic Sobolev spaces. *SIAM J. Math. Anal.*, 23(3): 799–819, 1992. pages 11

[204] N. Kechkar and D. J. Silvester. Analysis of locally stabilized mixed finite element methods for the Stokes problem. *Math. Comp.*, 58(197):1–10, 1992. pages 82

[205] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31(3):482–528, 1991. pages 307, 308, 311, 312

[206] S. N. Kružkov. First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)*, 81 (123):228–255, 1970. pages 274

[207] A. Kufner, O. John, and S. Fučík. *Function spaces*. Noordhoff International Publishing, Leyden, The Netherlands; Academia, Prague, Czechia, 1977. pages 93, 94, 95, 96

[208] J. Kuntzmann. Neuere Entwicklungen der Methode von Runge und Kutta. *ZAMM-Z. Angew. Math. Mech.*, 41(S1):T28–T31, 1961. pages 168

[209] D. Kuzmin, R. Löhner, and S. E. Turek. *Flux–Corrected Transport*. Scientific Computation. Springer, Berlin, Germany, 2005. pages 327

[210] M.-J. Lai and L. L. Schumaker. *Spline functions on triangulations*, volume 110 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, UK, 2007. pages 296, 299, 300

[211] P. D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, 7:159–193, 1954. pages 295, 300

[212] P. D. Lax. Hyperbolic systems of conservation laws. II. *Comm. Pure Appl. Math.*, 10: 537–566, 1957. pages 284, 290

[213] P. G. LeFloch. *Hyperbolic Systems of Conservation Laws: The theory of classical and nonclassical shock waves*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, Switzerland, 2002. pages 281

[214] P. Lesaint. *Sur la résolution des sytèmes hyperboliques du premier ordre par des méthodes d'éléments finis*. PhD thesis, University of Paris VI, 1975. pages 57

[215] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, WI, Academic Press, New York, NY, 1974. pages 57, 153

[216] A. Linke and L. G. Rebholz. Pressure-induced locking in mixed methods for time-dependent (Navier-)Stokes equations. *J. Comput. Phys.*, 388:350–356, 2019. pages 206

[217] J.-L. Lions. Quelques résultats d'existence dans des équations aux dérivées partielles non linéaires. *Bull. Soc. Math. France*, 87:245–273, 1959. pages 198

[218] J.-L. Lions. Quelques remarques sur les équations différentielles opérationnelles du $\mathbf{1^{er}}$ ordre. *Rend. Sem. Mat. Univ. Padova*, 33:213–225, 1963. pages 108, 181

[219] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*, volume 1. Dunod, Paris, France, 1969. pages 101, 198

[220] J.-L. Lions and E. Magenes. *Non-homogeneous Boundary Value Problems and Applications. Vols. I, II*. Springer-Verlag, New York-Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181-182. pages 100, 101, 108, 181

[221] T. P. Liu. The Riemann problem for general systems of conservation laws. *J. Differential Equations*, 18:218–234, 1975. pages 290

[222] A. J. Lučka. The rate of convergence to zero of the residual and the error for the Bubnov-Galerkin method and the method of least squares. In *Proc. Sem. Differential and Integral Equations, No. I (Russian)*, pages 113–122. Akad. Nauk Ukrain. SSR Inst. Mat., Kiev, Ukraine, 1969. pages 17

[223] C. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104(4):489–514, 2006. pages 154, 155

[224] M. Marion and R. Temam. Nonlinear Galerkin methods. *SIAM J. Numer. Anal.*, 26(5): 1139–1157, 1989. pages 117

[225] M. Marion and R. Temam. Nonlinear Galerkin methods: the finite elements case. *Numer. Math.*, 57(3):205–226, 1990. pages 117

[226] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.*, 41(4):713–742, 2007. pages 40, 41, 82

[227] G. Matthies, P. Skrzypacz, and L. Tobiska. Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.*, 32:90–105, 2008. pages 41, 62

[228] P. Monk and G. R. Richter. A discontinuous Galerkin method for linear symmetric hyperbolic systems in inhomogeneous media. *J. Sci. Comput.*, 22/23:443–477, 2005. pages 253

[229] M. Neumüller and I. Smears. Time-parallel iterative solvers for parabolic evolution equations. *SIAM J. Sci. Comput.*, 41(1):C28–C51, 2019. pages 187

[230] R. H. Nochetto and J.-H. Pyo. The gauge-Uzawa finite element method. I. The Navier-Stokes equations. *SIAM J. Numer. Anal.*, 43(3):1043–1068, 2005. pages 222

[231] A. Nouy. A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 199 (23-24):1603–1626, 2010. pages 162

[232] O. A. Oleĭnik. Discontinuous solutions of non-linear differential equations. *Uspehi Mat. Nauk (N.S.)*, 12(3(75)):3–73, 1957. pages 274

[233] O. A. Oleĭnik. Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation. *Uspehi Mat. Nauk*, 14(2 (86)):165–170, 1959. pages 274

[234] S. Osher. The Riemann problem for nonconvex scalar conservation laws and Hamilton-Jacobi equations. *Proc. Amer. Math. Soc.*, 89(4):641–646, 1983. pages 276

[235] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 19(5):816–829, 2012. pages 189

[236] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov. Least-Squares mixed finite elements for second-order elliptic problems. *SIAM J. Numer. Anal.*, 31(5):1368–1377, 1994. pages 28

[237] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1):119–130, 1996. pages 300

[238] R. Picard. A structural observation for linear material laws in classical mathematical physics. *Math. Methods Appl. Sci.*, 32(14):1768–1803, 2009. pages 236

[239] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Germany, second edition, 1997. pages 62

[240] R. Rannacher. On Chorin's projection method for the incompressible Navier-Stokes equations. In *The Navier-Stokes Equations II—Theory and Numerical Methods (Oberwolfach, Germany, 1991)*, volume 1530 of *Lecture Notes in Math.*, pages 167–183. Springer, Berlin, Germany, 1992. pages 217

[241] J. Rauch. Boundary value problems with nonuniformly characteristic boundary. *J. Math. Pures Appl. (9)*, 73(4):347–353, 1994. pages 8

[242] P.-A. Raviart. The use of numerical integration in finite element methods for solving parabolic equations. In *Topics in numerical analysis (Proc. Roy. Irish Acad. Conf., University Coll., Dublin, Ireland, 1972)*, pages 233–264. Academic Press, London, UK, 1973. pages 115

[243] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Germany, second edition, 2008. pages 20, 62

[244] S. J. Ruuth. Global optimization of explicit strong-stability-preserving Runge-Kutta methods. *Math. Comp.*, 75(253):183–207, 2006. pages 311

[245] S. J. Ruuth and R. J. Spiteri. High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations. *SIAM J. Numer. Anal.*, 42(3):974–996, 2004. pages 309

[246] F. Schieweck. On the role of boundary conditions for CIP stabilization of higher order finite elements. *Electron. Trans. Numer. Anal.*, 32:1–16, 2008. pages 62

[247] L. Schmutz and T. P. Wihler. The variable-order discontinuous Galerkin time stepping scheme for parabolic evolution problems is uniformly L^∞-stable. *SIAM J. Numer. Anal.*, 57 (1):293–319, 2019. pages 153

[248] D. Schötzau and C. Schwab. Time discretization of parabolic problems by the *hp*-version of the discontinuous Galerkin finite element method. *SIAM J. Numer. Anal.*, 38(3):837–875, 2000. pages 153

[249] D. Schötzau and T. P. Wihler. A posteriori error estimation for *hp*-version time-stepping methods for parabolic partial differential equations. *Numer. Math.*, 115(3):475–509, 2010. pages 153, 154, 167

[250] C. Schwab and R. Stevenson. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comp.*, 78(267):1293–1318, 2009. pages 181

[251] D. Serre. *Systems of conservation laws. 1. Hyperbolicity, Entropies, Shock Waves*. Cambridge University Press, Cambridge, UK, 1999. Translated from the 1996 French original by I. N. Sneddon. pages 284

[252] J. Shen. On pressure stabilization method and projection method for unsteady Navier-Stokes equations. In G. R. R. Vichnevetsky, D. Knight, editor, *Advances in computer methods for partial differential equations VII*, pages 658–662. IMACS, 1992. pages 217

[253] J. Shen. On error estimates of projection methods for the Navier-Stokes equations: first-order schemes. *SIAM J. Numer. Anal.*, 29:57–77, 1992. pages 217

[254] J. Shen. A remark on the projection-3 method. *Internat. J. Numer. Methods Fluids*, 16(3): 249–253, 1993. pages 218

[255] J. Shen. On error estimates of the penalty method for unsteady Navier-Stokes equations. *SIAM J. Numer. Anal.*, 32(2):386–403, 1995. pages 225, 227

[256] J. Shen. On error estimates of projection methods for the Navier-Stokes equations: second-order schemes. *Math. Comp.*, 65(215):1039–1065, 1996. pages 218

[257] R. E. Showalter. *Monotone operators in Banach spaces and nonlinear partial differential equations*, volume 49 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1996. pages 94, 236

[258] C.-W. Shu. Total-variation-diminishing time discretizations. *SIAM J. Sci. Statist. Comput.*, 9(6):1073–1084, 1988. pages 309

[259] C.-W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439 – 471, 1988. pages 308, 309

[260] D. Silvester. Optimal low order finite element methods for incompressible flow. *Comput. Methods Appl. Mech. Engrg.*, 111(3-4):357–368, 1994. pages 82

[261] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146:65–96, 1987. pages 101

[262] I. Smears. Robust and efficient preconditioners for the discontinuous Galerkin time-stepping method. *IMA J. Numer. Anal.*, 37(4):1961–1985, 2017. pages 154, 162, 187, 189

[263] J. Smoller. *Shock waves and reaction-diffusion equations*, volume 258 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York-Berlin, 1983. pages 291

[264] H. Sohr and W. von Wahl. On the regularity of the pressure of weak solutions of Navier-Stokes equations. *Arch. Math. (Basel)*, 46(5):428–439, 1986. pages 198

[265] R. J. Spiteri and S. J. Ruuth. A new class of optimal high-order strong-stability-preserving time discretization methods. *SIAM J. Numer. Anal.*, 40(2):469–491, 2002. pages 311

[266] J. C. Strikwerda and Y. S. Lee. The accuracy of the fractional step method. *SIAM J. Numer. Anal.*, 37(1):37–47, 1999. pages 218

[267] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984. pages 334

[268] E. Tadmor. Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comp.*, 43(168):369–381, 1984. pages 300

[269] E. Tadmor. Convergence of spectral methods for nonlinear conservation laws. *SIAM J. Numer. Anal.*, 26(1):30–44, 1989. pages 41

[270] F. Tantardini and A. Veeser. The L^2-projection and quasi-optimality of Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 54(1):317–340, 2016. pages 119, 181, 183, 184, 185

[271] R. Temam. Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires ii. *Arch. Rat. Mech. Anal.*, 33:377–385, 1969. pages 213, 216, 227

[272] R. Temam. *Navier–Stokes equations*, volume 2 of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, The Netherlands, 1977. pages 193

[273] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Germany, second edition, 2006. pages 115, 138, 145, 146, 153, 160

[274] T. Thompson. A discrete commutator theory for the consistency and phase error analysis of semi-discrete C^0 finite element approximations to the linear transport equation. *J. Comput. Appl. Math.*, 303:229–248, 2016. pages 324

[275] L. J. P. Timmermans, P. D. Minev, and F. N. van de Vosse. An approximate projection scheme for incompressible flow using spectral elements. *Internat. J. Numer. Methods Fluids*, 22:673–688, 1996. pages 219

[276] L. Tobiska and R. Verfürth. Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations. *SIAM J. Numer. Anal.*, 33(1):107–127, 1996. pages 75

[277] E. F. Toro. *Riemann solvers and numerical methods for fluid dynamics*. Springer-Verlag, Berlin, third edition, 2009. pages 290

[278] A. Toselli. *hp* discontinuous Galerkin approximations for the Stokes problem. *Math. Models Methods Appl. Sci.*, 12(11):1565–1597, 2002. pages 88

[279] K. Urban and A. T. Patera. A new error bound for reduced basis approximation of parabolic partial differential equations. *C. R. Math. Acad. Sci. Paris*, 350:203–207, 2012. pages 181

[280] J. van Kan. A second-order accurate pressure-correction scheme for viscous incompressible flow. *SIAM J. Sci. Statist. Comput.*, 7(3):870–891, 1986. pages 217

[281] N. N. Vladimirova, B. G. Kuznetsov, and N. N. Yanenko. Numerical calculation of the symmetrical flow of viscous incompressible liquid around a plate (in Russian). In *Some problems in computational and applied mathematics*. Nauka, Novosibirsk, 1966. pages 227

[282] M. Vlasák and F. Roskovec. On Runge-Kutta, collocation and discontinuous Galerkin methods: mutual connections and resulting consequences to the analysis. In *Programs and algorithms of numerical mathematics 17*, pages 231–236. Acad. Sci. Czech Repub. Inst. Math., Prague, 2015. pages 170

[283] E. L. Wachspress. *A rational finite element basis*. Academic Press, Inc. [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Mathematics in Science and Engineering, Vol. 114. pages 317

[284] J. Warren, S. Schaefer, A. N. Hirani, and M. Desbrun. Barycentric coordinates for convex sets. *Adv. Comput. Math.*, 27(3):319–338, 2007. pages 317

[285] M. F. Wheeler. A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations. *SIAM J. Numer. Anal.*, 10:723–759, 1973. pages 116

[286] T. P. Wihler. An a priori error analysis of the *hp*-version of the continuous Galerkin FEM for nonlinear initial value problems. *J. Sci. Comput.*, 25(3):523–549, 2005. pages 167

[287] N. N. Yanenko. *The method of fractional steps. The solution of problems of mathematical physics in several variables*. Springer-Verlag, New York, 1971. pages 227

[288] L. Ying. A second order explicit finite element scheme to multi-dimensional conservation laws and its convergence. *Sci. China Ser. A*, 43(9):945–957, 2000. pages 265

[289] K. Yosida. *Functional analysis*. Classics in Mathematics. Springer-Verlag, Berlin, Germany, 1995. Reprint of the sixth (1980) edition. pages 236, 238

[290] R. Young. The *p*-system. I. The Riemann problem. In *The legacy of the inverse scattering transform in applied mathematics (South Hadley, MA, 2001)*, volume 301 of *Contemp. Math.*, pages 219–234. Amer. Math. Soc., Providence, RI, 2002. pages 290, 292, 303

[291] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31(3):335–362, 1979. pages 327, 328

[292] Q. Zhang and C.-W. Shu. Error estimates to smooth solutions of Runge-Kutta discontinuous Galerkin methods for scalar conservation laws. *SIAM J. Numer. Anal.*, 42(2):641–666, 2004. pages 265

[293] Q. Zhang and C.-W. Shu. Stability analysis and a priori error estimates of the third order explicit Runge-Kutta discontinuous Galerkin method for scalar conservation laws. *SIAM J. Numer. Anal.*, 48(3):1038–1063, 2010. pages 266

[294] M. Zlámal. Finite element multistep discretizations of parabolic boundary value problems. *Math. Comp.*, 29:350–359, 1975. pages 138

Index